Editors' Suggestion

Statistical limits of dictionary learning: Random matrix theory and the spectral replica method

Jean Barbier [*]

International Center for Theoretical Physics (ICTP), I–34151 Trieste, Italy

Nicolas Macris [†]

Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

We consider increasingly complex models of *matrix denoising* and *dictionary learning* in the Bayes-optimal setting, in the challenging regime where the matrices to infer have a rank growing *linearly* with the system size. This is in contrast with most existing literature concerned with the low-rank (i.e., constant-rank) regime. We first consider a class of rotationally invariant matrix denoising problems whose mutual information and minimum mean-square error are computable using techniques from random matrix theory. Next, we analyze the more challenging models of dictionary learning. To do so we introduce a combination of the replica method from statistical mechanics together with random matrix theory, coined *spectral replica method*. This allows us to derive variational formulas for the mutual information between hidden representations and the noisy data of the dictionary learning problem, as well as for the overlaps quantifying the optimal reconstruction error. The proposed method reduces the number of degrees of freedom from $\Theta(N^2)$ matrix entries to $\Theta(N)$ eigenvalues (or singular values), and yields Coulomb gas representations of the mutual information which are reminiscent of matrix models in physics. The main ingredients are a combination of large deviation results for random matrices together with a replica symmetric decoupling ansatz at the level of the probability distributions of eigenvalues (or singular values) of certain overlap matrices and the use of Harish-Chandra-Itzykson-Zuber spherical integrals.

I. INTRODUCTION

The simplest linear-rank matrix inference task is the problem of recovering the rotationally invariant full-rank matrix S from noisy observations Y generated as

$$Y = \sqrt{\lambda}S + \xi,$$

where ξ is some Wigner Gaussian noise matrix. We refer to this problem as *matrix denoising*. In the random matrix theory (RMT) literature, typical problems are concerned with deriving spectral properties: spectral density and correlation functions of the eigenvalues or singular values of Y, its bulk statistics, the fluctuations of its largest and smallest eigenvalues, potential universality properties, etc. The literature is too large to be exhaustively reviewed here, and relevant references will be cited along the paper. We refer to [1,2] for mathematics books or [3–5] for more physics-oriented presentations. In this paper we instead consider information-theoretic questions such as "given a certain signal-to-noise ratio λ, what is the mutual-information between the hidden matrix signal S and the observed noisy data Y?" or "what is the statistically optimal reconstruction error on S?". We are interested in answering these questions in certain asymptotic large size limits. Despite the apparent simplicity of the model, these questions turn out to be highly nontrivial.

In matrix denoising we are "only" interested in the reconstruction of the matrix S. This allows us to analyze the model using solely RMT. But there exist models where S possesses some additional internal structure other than the (possibly nontrivial) statistics of its spectrum and/or S may not be rotationally invariant. This is the case in the model we study next: *dictionary learning*, where a product structure arises.

Let M noisy N-dimensional data points $(Y_j)_{j \leqslant M}$ be stacked as the columns of $Y \in \mathbb{K}^{N \times M}$, with $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. The unsupervised dictionary learning task we consider is to find a representation of this data Y in the form

$$ST^\dagger + Z.$$

The unknowns are both the "dictionary" $S \in \mathbb{K}^{N \times K}$ made of K features and the coefficients $T \in \mathbb{K}^{M \times K}$ in the decomposition of the clean data $ST^\dagger$ in feature basis. Here Z represents undesired noise. We also analyze a symmetric (or Hermitian) version of the problem where one aims to find a positive-definite first term in the decomposition of $Y \in \mathbb{K}^{N \times N}$ of the form

$$XX^\dagger + Z,$$

where $X \in \mathbb{K}^{N \times M}$. The rich internal structure coming from the product between matrices requires new ideas for the analysis: RMT alone does not seem sufficient for analyzing the optimal reconstruction performance on X, S, T themselves (instead of the products $ST^\dagger$, $XX^\dagger$ seen as individual matrices). This is

---

*Corresponding author: jbarbier@ictp.it
†nicolas.macris@epfl.ch

where the statistical mechanics of spin glasses, and in particular the *spectral replica method* developed in this paper, enter.

We will consider the Bayes-optimal "matched setting" where Y is truly generated according to one of the models described above, and the statistician has perfect knowledge of (1) this data-generating model (i.e., knows the additive nature and statistics of the noise and therefore the likelihood distribution) as well as (2) the prior distributions underlying the hidden random matrix signals X, S, T. The statistician can thus exploit this knowledge to write the correct posterior distribution in order to perform inference. Each model will be analyzed in both cases of real and complex matrices.

Given its fundamental nature and central role in signal processing and machine learning [6,7], dictionary learning has generated a large body of work with applications in representation learning [8], sparse coding [9–11], robust principal components analysis [12,13], submatrix localization [14], blind source separation [15], matrix completion [16,17], and community detection [18–20]. *Low-rank* (i.e., finite-rank) versions of dictionary learning have been introduced in statistics under the name of "spike models" as statistical models for sparse principal components analysis (PCA) [21–24]. In the low-rank regime $M, N \to +\infty$ proportionally and $K = \Theta(1)$ (or $N \to +\infty$ and $M = \Theta(1)$ for the symmetric case), these models have become paradigms for the study of phase transition phenomena in the recovery of low-rank information hidden in noise. In PCA classical rigorous results go back to Baik, Ben-Arous, and Péché [25,26] who analyzed the performance of spectral algorithms. More recently, low-dimensional variational formulas for the mutual information and corresponding phase transitions at the level of the Bayes-optimal minimum mean-square error estimator, as well as the algorithmic transitions of message passing and gradient descent-based algorithms and their associated computational-to-statistical gaps, have been derived due to the global effort of an highly interdisciplinary community [19,27–51].

In contrast, *much* less is known in the challenging *linear-rank* regime studied here where $N, M, K \to +\infty$ at similar rates, so that $\mathrm{rank}(ST^{\dagger})$ or $\mathrm{rank}(XX^{\dagger})$ diverge linearly with N. References closely related to our work are [52–56], which consider the same Bayesian setting. It has been suspected for some time and eventually confirmed, by the authors of the very recent study [57], that the mean field replica assumptions of these works give approximate formulas which are not asymptotically exact. Related problems affect the state evolution analysis [54,55] of message passing algorithms developed in these papers and in [58,59]. This state of affairs for linear-rank dictionary learning is to be contrasted to the low-rank situation where both the replica solutions and state evolution analyses of message passing algorithms are under rigorous mathematical control. We will discuss the differences between the replica assumptions in the above works and our spectral replica approach in a dedicated section. Another important work is [60] which considers the same models as ours but focuses on a certain class of rotational invariant estimators instead of the information-theoretic performance [61].

The main contribution of this paper is a *spectral replica ansatz* introduced in Secs. V and VI which seems to us the "minimal" possible ansatz to perform a replica symmetric calculation. This ansatz leads to variational expressions for the mutual information between hidden representations and noisy data, whose solutions yield overlaps quantifying the theoretical reconstruction error. Because the replica ansatz is minimal and the setting is Bayes-optimal we believe that our variational expressions are exact. Their solution, however, is far from trivial and is beyond the scope of this paper. We are still able to check the validity of our formulas analytically in a special case. In Secs. II, III, and IV we start by addressing the easier denoising problem for *rotation invariant* matrix ensembles. We propose expressions, conjectured to be exact, for the mutual information and mean-square errors. These are again challenging to explicitly compute as a function of the signal-to-noise ratio because they involve nontrivial *spherical integrals* [62,63]. But this time we are able to exploit Bayes optimality to get rid of the variational problem. Here we limit ourselves to provide perturbative expansions for some examples. We refer to [57] for recent work on the denoising problem, also providing expansions and numerical solutions.

Spherical integrals and the associated Harish-Chandra-Itzykson-Zuber formula [62,63] turn out to be an important tool in our analysis. In physics such integrals appear in multi-matrix models from high-energy physics with applications in string theory, quantum gravity, quantum chromodynamics, fluctuating surfaces, and map enumeration [64–75]. In certain aspects multimatrix models from high-energy physics are similar to matrix *inference* models appearing in denoising and dictionary learning. For example, in contrast to the low-dimensional order parameters of standard low-rank inference problems [76], in linear-rank regimes the order parameters are eigenvalue and singular value densities, as in multimatrix models. Also, at first sight, the variational formulas found in the present contribution look very much like those appearing in these models; see [77]. It may thus be tempting to think that matrix inference models are special cases of known matrix models. This is *not* the case, however. The presence of frozen, correlated randomness in inference, namely, the data, radically changes the nature of the problem: new tools are needed. This essential difference prevents borrowing various important techniques from this field, but nevertheless certain methods used in the analysis of matrix models will be crucial, in particular the use of spherical integrals in order to integrate rotational degrees of freedom.

Concretely, having frozen data Y translates, as in spin glasses [78,79], into the need to evaluate the expectation of the *logarithm* of the partition function with respect to it. This yields the behavior of the model for typical realizations of the signals and data. The difference is clear: the canonical two-matrix model from physics reads [66,68,77]

$$\ln \mathcal{Z}_{2MM} = \ln \int dA \, dB \exp \mathrm{Tr}[f(A) + g(B) + h(AB)],$$

with $A = A^{\dagger}$, $B = B^{\dagger}$ and with f, g, h just depending on the spectra. Instead the free energy of matrix inference models (which is essentially the Shannon entropy of the data) looks like

$$\mathbb{E}_Y \ln \mathcal{Z}_{INFER}(Y) = \int dY \exp[\mathrm{Tr} f(Y)]$$

$$\times \ln \int dU \exp \mathrm{Tr}[g(U) + h(UY)]. \quad (1)$$

This is a model with "quenched randomness" that needs to be averaged, namely, the data, while the log-partition function of the two-matrix model is nonrandom and is a purely "annealed" model in physics terminology. This form gets more complicated in nonsymmetric multipartite systems such as the $ST^\dagger$-dictionary learning problem, with an integration over more matrices which are not necessarily symmetric or Hermitian. The presence of a quenched average in (1) is far from innocent. It generates a whole new level of difficulty, since the methods in the previous references all rely on a direct saddle-point evaluation and do not apply [77]. This quenched average is the reason behind the fact that, even if only Hermitian matrices are present in the original inference model, non-Hermitian matrices will appear along the analysis. The role of the spectral replica method combined with RMT is precisely to deal with these new difficulties, at a nonrigorous level.

Many derivations presented in this paper are based on heuristics, yet they are conjectured to be exact in proper asymptotic limits. We believe that some of our methodology, in particular the spectral replica method, will pave the way to the analysis of a class of inference and learning problems involving large linear-rank matrices, which remained inaccessible until now. Moreover, given the breadth of applications of such disordered matrix models in information processing systems but also physics, our results may have an impact in a broader context.

A. Organization

In Sec. II we start by analyzing the simplest linear-rank matrix inference model using RMT techniques, namely, denoising of an Hermitian rotationally invariant matrix. Two special cases (one only being nontrivial) are completely treated, in the sense of deriving explicit enough formulas to draw a phase diagram. In Sec. III we provide generic systematic expansions of the previously derived formulas in the low signal-to-noise regimes. Section IV generalizes the analysis to the case of non-Hermitian matrix denoising. Section V is devoted to the analysis of Hermitian (i.e., positive-definite) dictionary learning. RMT tools do not suffice anymore, and we introduce the *spectral replica method* to go forward. The first part of this section reduces the model to effective Coulomb gases of singular values, while the second one expresses it in terms of eigenvalues. We also discuss the main differences with previous attempts to analyze this model. Finally in Sec. VI we consider the nonsymmetric case of dictionary learning. Appendix A recalls known facts about full-rank spherical integrals which are of crucial importance in our analyses. Appendix B derives a generic formula for the minimum mean-square error in matrix denoising. The last Appendix, Appendix C, provides the necessary MATHEMATICA codes to reproduce our numerical results.

B. Notations

Let the field $\mathbb{K} = \mathbb{R}$ if $\beta = 1$ or $\mathbb{K} = \mathbb{C}$ if $\beta = 2$, where β refers to the Dyson index. The symbol $\dagger$ corresponds to the transpose τ in the real case $\beta = 1$ or to the transpose conjugate when $\beta = 2$, the conjugate being $\bar{z} := \mathrm{Re}z - i\,\mathrm{Im}z$ with

Rez and Imz the real and imaginary parts of $z \in \mathbb{C}$ and $i = \sqrt{-1}$. Vectors and matrices are in bold. Vectors are columns by default and their transpose (conjugate) $x^\dagger$ are row vectors. When no confusion can arise we denote the trace $\mathrm{Tr}f(A) = \mathrm{Tr}[f(A)]$ so, e.g., $\mathrm{Tr}A^2 = \mathrm{Tr}[A^2]$ or $\mathrm{Tr}AB = \mathrm{Tr}[AB]$. Similarly $\mathbb{E}X^2 = \mathbb{E}(X)^2 = \mathbb{E}[X^2] \geqslant (\mathbb{E}X)^2 = \mathbb{E}[X]^2$. The usual inner product between (possibly complex) vectors $\sum_i \bar{x}_i y_i$ is denoted $x^\dagger y$ or $\langle x, y \rangle$ (with $\bar{x} = x$ if x is real); the matrix inner product is $\mathrm{Tr}X^\dagger Y$. The usual L^2 vector (squared) norm is $\sum_i |x_i|^2 = \|x\|^2$. Every sum or product over $j \leqslant t$ means over $j = 1, \ldots, t$. We will often drop parentheses, e.g., $\exp \cdots = \exp(\cdots)$. Depending on the context, the symbol $\propto$ means "equality up to a normalization," "equality up to an irrelevant additive constant," or "proportional to." We denote $[N] := \{1, \ldots, N\}$. For a diagonal matrix Σ we will write the diagonal elements with a single index $\Sigma_i := \Sigma_{ii}$. For a diagonalizable matrix A the diagonal matrix of eigenvalues is $\lambda^A = \lambda_A$ and the individual eigenvalues are λ_i^A; similarly for a generic matrix B the matrix of singular values is $\sigma^B = \sigma_B$ (with the singular values on the main diagonal) and the individual singular values are σ_i^B. We generically denote ρ_A the asymptotic limit of the empirical distribution of eigenvalues or singular values of a matrix A. Symbol $\mathcal{P}$ refers to the set of probability densities with finite support; $\mathcal{P}_{\geqslant 0}$ emphasizes that the support is nonnegative. A matrix M with elements M_{ij} may also be written $[M_{ij}]$ or $[M_{cd}]$. Finally, the symbol $x \sim y$ means equality in distribution for two random variables; $x \sim p$ instead means that x is a sample from p whenever p is a probability distribution or a sample from $p(x)dx$ if p is a probability density function.

II. DENOISING OF AN HERMITIAN ROTATIONALLY INVARIANT MATRIX

We start with the simplest possible model of inference of a large matrix: linear-rank rotationally invariant matrix denoising. It will require only known tools from random matrix theory.

A. The model

Let a matrix signal $S = S^\dagger \in \mathbb{K}^{N \times N}$ with $S \sim P_{S,N}$ for some known prior distribution (which in general does not factorize over the matrix entries), and $\xi = \xi^\dagger \in \mathbb{K}^{N \times N}$ a standard Wigner noise matrix with probability density function (p.d.f.)

$$dP_{\xi,N}(\xi) = C_N\, d\xi \exp \mathrm{Tr}\left(-\frac{\beta N}{4}\xi^2\right)$$

with C_N the normalization factor. Consider a matrix denoising problem with data $Y = Y^\dagger \in \mathbb{K}^{N \times N}$ generated according to the observation model

$$Y = \sqrt{\lambda}S + \xi. \tag{2}$$

The hidden matrix S to be recovered from the data is rotationally invariant in the sense that it is drawn from a prior distribution such that

$$dP_{S,N}(S) = dP_{S,N}(O^\dagger SO)$$

for any orthogonal ($\beta = 1$) or unitary ($\beta = 2$) matrix O. It can thus be diagonalized as $S = \tilde{U}^\dagger \lambda^S \tilde{U}$ where $\tilde{U} \sim \mu_N^{(\beta)}$ with

$\mu_N^{(\beta)}$ the normalized Haar measure over the orthogonal group $\mathcal{O}(N)$ if $\beta = 1$ or over the unitary group $\mathcal{U}(N)$ if $\beta = 2$. Matrix S has $O(1/\sqrt{N})$ entries. This scaling for the entries of S and of the Wigner matrix are such that the (real) eigenvalues of S, ξ and therefore Y remain $O(1)$ in the limit $N \to +\infty$. The joint probability density function (j.p.d.f.) of eigenvalues of the matrix Y generated according to model (2) is rigorously established in the case where S has independent entries (which we do *not* necessarily assume), and is obtained with techniques of a similar flavor as our strategy (i.e., based on the use of spherical integrals) [80]; see also [1] for an approach based on Dyson's Brownian motion.

The above model defines a random matrix ensemble for Y which is linked to the Rosenzweig-Porter random matrix model [81] from condensed matter. A generalized version of it has a rich behavior with a localization transition and regions with "multifractal eigenstates"; see [82–85]. The regime we are interested in, namely, with both λ^S and the eigenvalues of ξ being order 1, corresponds precisely to the critical scaling regime where a recently discovered transition from nonergodic extended states to ergodic extended states happens in the model (i.e., the transition towards multifractality; see the discussion on the regime $\gamma = 1$ in [84]). We find the connection with inference particularly intriguing and the results of this paper may thus may be of independent interest in the condensed matter context. In particular, if there is an information-theoretic transition in the inference problem it may happen to be connected to the $\gamma = 1$ ergodicity-breaking transition found in [84].

We consider a generic j.p.d.f. $p_{S,N}(\lambda^S)$ of eigenvalues which is symmetric [86] (i.e., invariant under any permutation of the entries of λ^S) and whose one-point marginal is assumed to weakly converge as $N \to +\infty$ to a well-defined measure ρ_S with finite support and without point masses; in particular S needs to be full-rank. We discuss at the end of this section how to overcome this latter constraint. Generically, rotational invariance implies that the prior over $S = \tilde{U}^\dagger \lambda^S \tilde{U}$ can be decomposed as

$$dP_{S,N}(S) = d\mu_N^{(\beta)}(\tilde{U}) \, dp_{S,N}(\lambda^S). \tag{3}$$

A (rather generic) special case of rotationally invariant measures for the signal are those of the form

$$dP_{S,N}(S) \propto \mu_N^{(\beta)}(\tilde{U}) \, d\lambda^S \exp \mathrm{Tr}\left[-\frac{\beta N}{4} V(\lambda^S) \right]$$
$$\times |\Delta_N(\lambda^S)|^\beta R \tag{4}$$

for a rotation invariant matrix potential $\mathrm{Tr}V(S) = \mathrm{Tr}V(\lambda^S)$, and where the Vandermonde determinant for a $N \times N$ diagonal matrix A with diagonal entries $(A_i)_{i \leqslant N}$ reads

$$\Delta_N(A) := \prod_{i<j}^{1,N} (A_i - A_j) = \det[(A_c)^{d-1}]. \tag{5}$$

In this case the eigenvalues j.p.d.f. has the form

$$p_{S,N}(\lambda^S) \propto \exp \mathrm{Tr}\left[-\frac{\beta N}{4} V(\lambda^S) \right] |\Delta_N(\lambda^S)|^\beta. \tag{6}$$

The case of a standard real symmetric or complex Hermitian Wigner matrices then corresponds to $V(\lambda^S) = \lambda_S^2$.

Wishart matrices $S = XX^\dagger$ with $X \in \mathbb{K}^{N \times M}$ have a density for $N \leqslant M$ corresponding to $V(\lambda^S) = 2[1 - M/N - 1/N + 2/(\beta N)] \ln \lambda^S + 2(M/N)\lambda^S$ (see more details later in Sec. III).

The main object of interest is the mutual information between data and signal:

$$I(Y; S) = H(Y) - H(Y|S)$$
$$= H(Y) - H(\xi)$$
$$= -\mathbb{E}_Y \ln \int dP_{S,N}(s) \, C_N \exp \mathrm{Tr}\left[-\frac{\beta N}{4}(Y - \sqrt{\lambda}s)^2 \right]$$
$$+ \mathbb{E} \ln C_N \exp \mathrm{Tr}\left(-\frac{\beta N}{4}\xi^2 \right)$$
$$= -\mathbb{E}_Y \ln \int dP_{S,N}(s) \exp \frac{\beta N}{2} \mathrm{Tr}\left(\sqrt{\lambda}sY - \frac{\lambda}{2}s^2 \right)$$
$$+ \frac{\beta \lambda N}{4} \mathbb{E} \mathrm{Tr}S^2. \tag{7}$$

B. Free entropy and mutual information through random matrix theory

We define the free entropy $f_N = f_N(Y)$ as minus the first term in (7) divided by N^2, without the expectation. The mutual information will directly be deduced from the free entropy, and using that the later is expected to concentrate onto its Y average. Using the eigendecomposition $s = U^\dagger \lambda^s U$ the free entropy reads

$$f_N := \frac{1}{N^2} \ln \int dP_{S,N}(s) \exp \frac{\beta N}{2} \mathrm{Tr}\left(\sqrt{\lambda}sY - \frac{\lambda}{2}s^2 \right)$$
$$= \frac{1}{N^2} \ln \int dp_{S,N}(\lambda^s) \exp \left(-\frac{\beta \lambda N}{4} \mathrm{Tr}\lambda_s^2 \right)$$
$$\times \int d\mu_N^{(\beta)}(U) \exp \frac{\beta \sqrt{\lambda}}{2} N \mathrm{Tr}(U^\dagger \lambda^s UY)$$
$$= \frac{1}{N^2} \ln \int d\lambda^s \exp N^2 \left[\frac{1}{N^2} \ln p_{S,N}(\lambda^s) \right.$$
$$\left. -\frac{\beta \lambda}{4N} \mathrm{Tr}\lambda_s^2 + I_N^{(\beta)}(\lambda^s, \lambda^Y, \sqrt{\lambda}) \right]. \tag{8}$$

There appears the Harish-Chandra-Itzykson-Zuber (HCIZ) spherical integral [63,87], which is only a function of the eigenspectra of its arguments. For $N \times N$ symmetric/Hermitian matrices A and B,

$$I_N^{(\beta)}(A, B, \gamma) = I_N^{(\beta)}(\lambda^A, \lambda^B, \gamma)$$
$$:= \frac{1}{N^2} \ln \int d\mu_N^{(\beta)}(U) \exp \frac{\beta \gamma}{2} N \mathrm{Tr}\,(U^\dagger \lambda^A U \lambda^B), \tag{9}$$

where the integration is over $\mathcal{O}(N)$ when $\beta = 1$ or $\mathcal{U}(N)$ when $\beta = 2$. We recall known facts about it in Appendix A. It has a well-defined limit [87,88]:

$$I^{(\beta)}[\rho_A, \rho_B, \gamma] := \lim_{N \to +\infty} I_N^{(\beta)}(\lambda^A, \lambda^B, \gamma), \tag{10}$$

where ρ_A, ρ_B are the asymptotic *densities* of eigenvalues (i.e., one-point correlation functions) of A and B, respectively. In the present case, the eigenvalues λ^Y of the data matrix Y and

associated asymptotic density ρ_Y are fixed by the model; $\boldsymbol{\lambda}^Y$ can be simulated and ρ_Y can be obtained using free probability; see, e.g., [89,90]. Thus, a standard saddle-point argument leads to the following conjecture for the free entropy $f_N = f_N(\boldsymbol{Y})$ as $N \to +\infty$.

Result 1 (Free entropy of Hermitian rotationally invariant matrix denoising). The free entropy of model (2) is

$$f_N = \sup_{\boldsymbol{\lambda}^s \in \mathbb{R}^N} \left[\frac{1}{N^2} \ln p_{S,N}(\boldsymbol{\lambda}^s) - \frac{\beta\lambda}{4N} \mathrm{Tr}\boldsymbol{\lambda}_s^2 + I_N^{(\beta)}(\boldsymbol{\lambda}^s, \boldsymbol{\lambda}^Y, \sqrt{\lambda}) \right] + \tau_N. \quad (11)$$

The constant τ_N fixes the constraint $f_N(\lambda = 0) = 0$ [the spherical integral (9) vanishes when $\lambda = 0$]:

$$\tau_N := - \sup_{\boldsymbol{\lambda}^s \in \mathbb{R}^N} \frac{1}{N^2} \ln p_{S,N}(\boldsymbol{\lambda}^s) + o_N(1).$$

Because we do not rigorously control the saddle-point estimation we state the result as a conjecture, but it should not be out of reach to turn it into a theorem using techniques as in [77]. This free entropy was not averaged with respect to $\boldsymbol{Y}$. But it is expected that additionally it is self-averaging as it depends on $\boldsymbol{Y}$ only through its spectrum $\boldsymbol{\lambda}^Y$:

$$\mathbb{E} f_N = f_N + o_N(1).$$

Note that from this conjecture, the minimum mean-square error (MMSE) can be deduced using the I-MMSE relation for Gaussian channels [91,92]

$$\frac{1}{N^2} \mathbb{E} \|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2$$
$$= \frac{4}{\beta N^2} \frac{d}{d\lambda} I(\boldsymbol{Y}; \boldsymbol{S}) + O(1/N)$$
$$= \frac{1}{N} \mathbb{E}\mathrm{Tr}\boldsymbol{\lambda}_S^2 - \frac{4}{\beta} \frac{d}{d\lambda} \mathbb{E} f_N + O(1/N). \quad (12)$$

Regularity of eigenvalues and singular values distributions. All through the paper we assume that all eigenvalues (and later singular values) distributions are such that empirical distributions of eigenvalues or singular values converge weakly to well-defined asymptotic probability densities with (1) (possibly disconnected) finite support and (2) without any point masses. Cases of distributions with point masses (such as a matrix $\boldsymbol{S}$ of rank lower than N with a point mass δ_0 in its eigenvalues distribution; for example, a rank-deficient Wishart matrix with $N > M$) can be approximated by considering regularizations. If the original signal matrix, say, $\boldsymbol{S} \in \mathbb{R}^{N \times N}$, has a rank$(\boldsymbol{S}) < N$ with a finite fraction of eigenvalues or singular values strictly null, one may instead consider from the beginning the same inference model but with full-rank signal $\boldsymbol{S}_\varepsilon := \boldsymbol{S} + \boldsymbol{Z}_\varepsilon$ where $\boldsymbol{Z}_\varepsilon$ is an independent rotationally invariant regularization with norm smaller than ε, such as a Wigner matrix with sufficiently small variance. In certain cases it should then be possible to obtain the j.p.d.f. of the resulting matrix ensemble. The asymptotic formulas for the free entropies and mutual informations are expected to be continuous in ε. Thus assuming that the convergence to the asymptotic value is uniform in ε, we can permute the

$N \to +\infty$ and $\varepsilon \to 0_+$ limits to obtain the formulas for "non-full-rank" cases and densities with point masses. For the rest of the paper, we will thus restrict all theoretical arguments to full-rank cases without point masses.

Let us also mention that despite we focus on full-rank square models of matrix denoising (2) with $\boldsymbol{S}$ a $N \times N$ matrix, we believe that by combining our approach together with the idea of "quadratization of rectangular matrices" found in [93], and exploited, e.g., in [94,95], then it should not require too much work to generalize the results on matrix denoising of Secs. II and III to the rectangular setting $\boldsymbol{S} \in \mathbb{K}^{N \times M}, N \neq M$.

C. Expressing the result using a density order parameter due to a large deviation principle

For the typical form of eigenvalues density (6) we have

$$\frac{\ln p_{S,N}(\boldsymbol{\lambda}^s)}{N^2} = \frac{\beta}{2N^2} \sum_{i \neq j}^{1,N} \ln \left| \lambda_i^s - \lambda_j^s \right| - \frac{\beta}{4N} \mathrm{Tr} V(\boldsymbol{\lambda}^s). \quad (13)$$

When potential V acts entrywise this can be expressed using the empirical spectral distribution (ESD)

$$\hat{\rho}_N^\lambda(x) := \frac{1}{N} \sum_{i \leqslant N} \delta(\lambda_i^s - x) \quad (14)$$

as follows:

$$\frac{1}{N^2} \ln p_{S,N}(\boldsymbol{\lambda}^s) = \frac{\beta}{2} \int d\hat{\rho}_N^\lambda(x) d\hat{\rho}_N^\lambda(y) \ln |x - y|$$
$$- \frac{\beta}{4} \int d\hat{\rho}_N^\lambda(x) V(x) + c_N.$$

The "constant" c_N is formally infinity and takes care of removing the divergence coming from the double integral when $x = y$; it simplifies with the same constant appearing in τ_N. This combined with (10) suggests that the formula (11) can be expressed as an optimization over a density of eigenvalues. To see that, note that the integral (8) can be written as

$$\frac{1}{N^2} \ln \int D\hat{p}_{S,N}[\hat{\rho}] \exp N^2 \left(-\frac{\beta\lambda}{4} \int d\hat{\rho}(x) x^2 + I^{(\beta)}[\hat{\rho}, \rho_Y, \sqrt{\lambda}] + o_N(1) \right), \quad (15)$$

where the density of the ESD induced by $p_{S,N}(\boldsymbol{\lambda}^s)$ can be formally written as

$$D\hat{p}_{S,N}[\hat{\rho}] = D[\hat{\rho}] \int dp_{S,N}(\boldsymbol{\lambda}^s) \delta(\hat{\rho}_N^\lambda - \hat{\rho}). \quad (16)$$

For a large class of potentials V a large deviation principle (LDP) at scale N^2 with rate functional L holds for the random ESD [96–98]:

$$D\hat{p}_{S,N}[\hat{\rho}] \sim D[\hat{\rho}] \exp(-N^2 L[\hat{\rho}]). \quad (17)$$

If L is known, plugging this expression in (15) and evaluating the integral by saddle point approximation over the density provides the desired expression.

Let us discuss how to obtain the rate function L. Given a potential V defining a rotationally invariant ensemble of symmetric or Hermitian matrices given by (3) or (4), the

associated rate function L of the ESD $\hat{\rho}_N^{\lambda}$ is rather generically given, up to an irrelevant additive constant, by [96–98]

$$L[\hat{\rho}] = \frac{\beta}{2} \int d\hat{\rho}(x) \left(\frac{V(x)}{2} - \int d\hat{\rho}(y) \ln|x-y| \right). \quad (18)$$

A direct way to see that starting from (16) is by using (13) to get

$$\int dp_{S,N}(\boldsymbol{\lambda}^s) \delta(\hat{\rho}_N^{\lambda} - \hat{\rho})$$

$$= \exp N^2 \left[\frac{\beta}{2} \int d\hat{\rho}(x) \, d\hat{\rho}(y) \ln|x-y| \right.$$

$$\left. - \frac{\beta}{4} \int d\hat{\rho}(x) V(x) + \frac{1}{N^2} \ln \int d\boldsymbol{\lambda}^s \delta(\hat{\rho}_N^{\lambda} - \hat{\rho}) \right]. \quad (19)$$

Now, the entropic contribution $H[\hat{\rho}] := \ln \int d\boldsymbol{\lambda}^s \delta(\hat{\rho}_N^{\lambda} - \hat{\rho})$ can be evaluated by introducing a Fourier representation of the Dirac delta:

$$H[\hat{\rho}] = \ln \int d\boldsymbol{\lambda}^s \int D[g]$$

$$\times \exp \left\{ i \int dx \, g(x) \left[N\hat{\rho}(x) - \sum_{i \leqslant N} \delta(\lambda_i^s - x) \right] \right\}$$

$$\propto -N \int d\hat{\rho}(x) \ln \hat{\rho}(x) = O(N).$$

We recognize N times the Shannon entropy of $\hat{\rho}$ [see Sec. 4.2 in Chapter 4, Eqs. (4.14)–(4.18) in [5], or Appendix C of [99]]. Therefore, at leading order $\exp \Theta(N^2)$ this entropic contribution is negligible, and we identify from (17) the rate functional (18) from the LDP approach.

Now imagine that the potential V underlying the RMT ensemble (4) is not known but, instead, its asymptotic spectral density is. The generalized Tricomi formula (see, e.g., Chapter 5 of [5]) states that for a rotationally invariant ensemble of symmetric or Hermitian random matrices whose j.p.d.f. of the matrix entries is (3), the equilibrium spectral density ρ describing its $O(1)$ eigenvalues verifies

$$\mathcal{H}_{\rho}(x) := \mathrm{Pr} \int d\rho(t) \frac{1}{x-t} = \frac{\beta}{4} V'(x), \quad (20)$$

where $\mathcal{H}_{\rho}$ is the Hilbert transform of density ρ whose domain is $\mathbb{R}$ and Pr is the principal value. When unknown, it allows us to reconstruct the potential V given ρ by integration:

$$V(x) = \frac{4}{\beta} \int^x \mathcal{H}_{\rho}(t) \, dt + C \quad (21)$$

for a certain irrelevant constant C that can be set to 0 without loss of generality. Therefore at the level of rigor of this paper, a RMT ensemble of the form (4) can equivalently be defined from the knowledge of the potential or its asymptotic spectral distribution.

With all these results in hand we can now restate the previous replica symmetric formula using a density order parameter:

Result 2 (Free entropy of Hermitian rotationally invariant matrix denoising). The asymptotic free entropy of model (2) when the signal is drawn from the random matrix ensemble

(4) is

$$f_N \to \sup_{\rho_s \in \mathcal{P}} \left[\frac{\beta}{2} \int d\rho_s(x) \, d\rho_s(y) \ln|x-y| - \frac{\beta\lambda}{4} \int d\rho_s(x) x^2 \right.$$

$$\left. - \frac{\beta}{4} \int d\rho_s(x) V(x) + I^{(\beta)}[\rho_s, \rho_Y, \sqrt{\lambda}] \right] + \tau.$$

The optimization is over the set $\mathcal{P}$ of probablity densities with finite support. The constant τ fixes the constraint $f_N(\lambda = 0) = 0$ and is given by

$$- \sup_{\rho_s \in \mathcal{P}} \left[\frac{\beta}{2} \int d\rho_s(x) \, d\rho_s(y) \ln|x-y| - \frac{\beta}{4} \int d\rho_s(x) V(x) \right].$$

D. Simplifications in the Bayes-optimal setting using the Nishimori identity

The above conjectures have already reduced the computation of an integral over $\Theta(N^2)$ degrees of freedom (the matrix elements) onto an optimization problem over $\Theta(N)$ eigenvalues or a functional optimization over a density. But we claim that because we are in the Bayesian optimal setting the formulas can be further simplified in the form of Result 3 below. In the matched setting the posterior is the "correct" one, and as a consequence, a fundamental property known as the Nishimori identity holds. This identity states that for any well-behaved function $g : \mathbb{R}^{N \times N} \mapsto \mathbb{R}$ we have (here we state a restricted form of the most general identity found in [76])

$$\mathbb{E} \langle g(\boldsymbol{s}) \rangle = \mathbb{E} \, g(\boldsymbol{S}), \quad (22)$$

where the signal $\boldsymbol{S} \sim P_{S,N}$, while $\boldsymbol{s}$ is a sample from the Bayes-optimal posterior j.p.d.f.

$$dP_{S|Y,N}(\boldsymbol{s}|\boldsymbol{Y})$$

$$= \frac{1}{\mathcal{Z}(\boldsymbol{Y})} dP_{S,N}(\boldsymbol{s}) \exp \mathrm{Tr} \left[-\frac{\beta N}{4} (\boldsymbol{Y} - \sqrt{\lambda}\boldsymbol{s})^2 \right],$$

and the Gibbs bracket $\langle \cdot \rangle$ is the associated expectation. In particular we have

$$\mathbb{E} \left\langle \frac{\mathrm{Tr} \boldsymbol{\lambda}_s^k}{N} \right\rangle = \mathbb{E} \frac{\mathrm{Tr} \boldsymbol{\lambda}_S^k}{N}, \quad (23)$$

for the kth moment of the empirical density of eigenvalues of the signal.

We now give an heuristic argument based on four steps and leading to Result 3 below, and believe that this may be the starting point of a rigorous proof strategy. Recall that $\hat{\rho}_N^{\lambda}$ given by (14) is the ESD associated with the eigenvalues $\boldsymbol{\lambda}^s$ of the posterior sample. Define its moments

$$m_{k,N} := \int dx \, x^k \hat{\rho}_N^{\lambda}(x) = \frac{\mathrm{Tr} \boldsymbol{\lambda}_s^k}{N}.$$

(1) First, note that in expression (8) or (15) the ESD plays the role of an order parameter for a "mean-field" free entropy functional given by the logarithm of the integrand in (15), or equivalently, by the functional to be extremized in Result 2. Concretely, one can express the integrand in (8) or (15) entirely in terms of the moments $(m_{k,N})_{k \geqslant 1}$.

(2) Second, we assume that the ESD $\hat{\rho}_N^*$ corresponding to the extremizer in Result 1 is such that its moments $m_{k,N}^*$

are close to the Gibbs averages $\langle m_{k,N}\rangle = N^{-1}\langle \mathrm{Tr}\boldsymbol{\lambda}_s^k\rangle$. In other words

$$m_{k,N}^* = \langle m_{k,N}\rangle + o_N(1).$$

This is a natural self-consistency hypothesis for any "replica-symmetric" mean-field theory, where the optimal value of the order parameter generally coincides with the Gibbs average (the reader may recall the solution of the Curie-Weiss model for the prime example of this mechanism). Replica symmetry, namely, the self-averaging or concentration of the order parameters (or the moments $(m_{k,N})_k$), is generically rigorously valid in Bayes-optimal inference of low-rank models [76,100] and we think that this property extends to linear-rank regimes.

(3) Third, we assume that the Gibbs expectation of the moments concentrates with respect to the data $\boldsymbol{Y}$: $\langle m_{k,N}\rangle = \mathbb{E}\langle m_{k,N}\rangle + o_N(1)$. This translates to

$$\langle m_{k,N}\rangle = N^{-1}\mathbb{E}\langle \mathrm{Tr}\boldsymbol{\lambda}_s^k\rangle + o_N(1)$$
$$= \mathbb{E}\langle (\lambda_1^s)^k\rangle + o_N(1).$$

This is again true in low-rank Bayes-optimal inference [76,100].

(4) Finally, from the two previous points and the Nishimori identity (23) we conclude

$$m_{k,N}^* = \mathbb{E}(\lambda_1^S)^k + o_N(1).$$

We have thus found that, somewhat remarkably, the extremizer in Result 1 has a corresponding ESD matching the one of the signal $\hat{\rho}_N^* = \hat{\rho}_{S,N}$. Taking $N \to +\infty$, the argument becomes exact: the empirical density $\hat{\rho}_N^* \to \rho^* = \rho_S$ the asymptotic density of eigenvalues of the signal. This yields the formula (24) below. In particular the supremum in both the non trivial term where the spherical integral appears in Result 1 and the constant term τ_N (or τ in the case of infinite N) are the same. Consequently we obtain a greatly simplified expression for the mutual information using relation (7), the fact that $N^{-1}\mathrm{Tr}\boldsymbol{S}^2 = N^{-1}\mathrm{Tr}\boldsymbol{\lambda}_S^2$ concentrates onto $\mathbb{E}(S_1)^2$ when $N \to +\infty$, and the concentration assumption for the free entropy $\mathbb{E}f_N = f_N + o_N(1)$. We also obtain a formula for the MMSE using the I-MMSE relation (12).

Result 3 (Mutual information of Hermitian rotationally invariant matrix denoising). Let $\boldsymbol{\lambda}^s \in \mathbb{R}^N$ be the eigenvalues of a random matrix $s \sim P_{S,N}$, i.e., $\boldsymbol{\lambda}^s \sim p_{S,N}$. The mutual information of model (2) verifies

$$\frac{1}{N^2}I(\boldsymbol{Y};\boldsymbol{S}) = \frac{\beta\lambda}{2N}\mathrm{Tr}\boldsymbol{\lambda}_s^2 - I_N^{(\beta)}(\boldsymbol{\lambda}^s,\boldsymbol{\lambda}^Y,\sqrt{\lambda}) + o_N(1).$$

Introducing $\rho_s = \rho_S$ the asymptotic spectral density of $s \sim P_{S,N}$ we get for $N \to +\infty$

$$\frac{1}{N^2}I(\boldsymbol{Y};\boldsymbol{S}) \to \frac{\beta\lambda}{2}\int d\rho_s(x)x^2 - I^{(\beta)}[\rho_s,\rho_Y,\sqrt{\lambda}]. \quad (24)$$

We deduce from (12) and a convexity argument (as $[I(\boldsymbol{Y};\boldsymbol{S})]_N$ is a sequence of concave functions in λ) that the minimum mean-square error verifies

$$\frac{1}{N^2}\mathbb{E}\|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2 = \frac{2}{N}\mathbb{E}\mathrm{Tr}\boldsymbol{\lambda}_s^2 - \frac{4}{\beta}\frac{d}{d\lambda}I_N^{(\beta)}$$
$$\times (\boldsymbol{\lambda}^s,\boldsymbol{\lambda}^Y,\sqrt{\lambda}) + o_N(1), \quad (25)$$

or, working with the eigenvalue densities, for $N \to +\infty$

$$\frac{1}{N^2}\mathbb{E}\|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2$$
$$\to 2\int d\rho_s(x)x^2 - \frac{4}{\beta}\frac{d}{d\lambda}I^{(\beta)}[\rho_s,\rho_Y,\sqrt{\lambda}].$$

Remark 1. The spherical integral $I^{(\beta)}$ or $I_N^{(\beta)}$ is difficult to compute. One route is to try using the HCIZ formula, but it is known that the ratio of determinants involved in the formula (see Appendix A) is notoriously difficult to evaluate analytically or even numerically. Another one is to employ its asymptotic hydrodynamic description [87,88], but this is challenging too. The HCIZ formula can be evaluated exactly in very special cases (e.g., the uniform and Wigner cases below) or perturbatively (see Sec. III), or approximated by using sampling techniques [101]. We wish to point out that an easy and nice application of the HCIZ formula is to check that the asymptotic mutual information obtained in Result 3 is the same for a signal $\boldsymbol{S}$ or its centered (traceless) version $\boldsymbol{S} - \mathrm{I}_{\mathrm{d},N}N^{-1}\mathrm{Tr}\boldsymbol{S}$ (where $\mathrm{I}_{\mathrm{d},N}$ is the identity of size N); this can be checked using basic properties of determinants. We know *a priori* that this should be so because in the Bayesian-optimal case the statistician knows the asymptotic value of $N^{-1}\mathrm{Tr}\boldsymbol{S} \to \mathbb{E}\lambda_1^S$ (which is nothing else than the first moment of distribution of the signal) and can subtract it from the data matrix, so information-theoretically it has no influence.

Remark 2. Result 3 was first obtained in the thesis of Schmidt [56] in the real case $\beta = 1$ (see Appendix 7). But what we believe are crucial steps and justifications were omitted in his derivation, and it is not obvious to us how the final (correct) result was obtained. In particular, [56] jumps from Eq. (8) to the final Result 3 without justification [see the transition from Eq. (A.79) to (A.83) in Appendix 7 of [56]].

Remark 3. The formula for the MMSE involves the derivative of the HCIZ formula with respect to λ. This can be computed in cases where some expression for the asymptotic value $I^{(\beta)}[\rho_s,\rho_Y,\sqrt{\lambda}]$ is known. This is for example the case in the sanity checks of the next paragraphs, and also in terms of perturbative expansions presented in Sec. III for small and large signal-to-noise ratio. It is possible to deduce from the HCIZ formula an expression for the derivative directly in terms of the eigenvalues and eigenvectors of $\boldsymbol{Y}$. While this is not directly used in the present paper it could be of interest in numerical approaches and for the analysis of the various variational problems in this paper. For this reason we include it in Appendix B.

Let us comment on an *a priori quite* surprising observation. Consider three scenarios for the prior over the eigenvalues $\boldsymbol{\lambda}^S$ of the signal:

(1) The prior over the eigenvalues is of the form (6). When $N \to +\infty$ (which can be thought of as a vanishing temperature limit), strongly coupled eigenvalues $\boldsymbol{\lambda}^S$ drawn according to $p_{S,N}(\boldsymbol{\lambda}^S)$ "freeze" into configurations of low energy (which includes the external potential V plus the long range Coulomb repulsion due to the Vandermonde). The resulting one-point marginal is a nontrivial density ρ_S.

(2) The prior is factorized as $p_{S,N}(\boldsymbol{\lambda}^S) = \prod_{i\leqslant N}\rho_S(\lambda_i^S)$, where ρ_S corresponds to the asymptotic marginal from the prior in case (a). In this case the prior does not induce any

sort of interaction among eigenvalues and fluctuations survive even in the limit $N \to +\infty$: the "temperature remains finite" and no "freezing" occurs.

(3) The eigenvalues are deterministic and given to the statistician, i.e., $p_{S,N}(\boldsymbol{\lambda}^S) = \delta(\boldsymbol{\lambda}^S - \boldsymbol{\lambda}_0^S)$, where the fixed configuration $\boldsymbol{\lambda}_0^S$ has an empirical density weakly converging to ρ_S.

By construction these three priors have the same one-point marginals (in the large size limit). For example, in the Wigner case, (1) would correspond to (6) with $V(\lambda^S) = \lambda_S^2$, and case (2) to $p_{S,N}(\boldsymbol{\lambda}^S) = \prod_{i \leqslant N} (4 - (\lambda_i^S)^2)^{1/2}/(2\pi)$ a product of semicircle laws. For case (3) one can generate a typical sample from priors (1) or (2) and fix it. For the Wishart ensemble it would correspond to $V(\lambda^S) = 2[1 - M/N - 1/N + 2/(\beta N)] \ln \lambda^S + 2(M/N)\lambda^S$ in (6) for case (A), and $p_{S,N}(\boldsymbol{\lambda}^S) = \prod_{i \leqslant N} \rho_{\mathrm{MP}}(\lambda_i^S)$ a product of Marcenko-Pastur laws (42).

Now, we claim that in all three cases Result 3 holds without any difference apart from possible $o_N(1)$ corrections. Indeed, in case (3) the integration over $\boldsymbol{\lambda}^s$ in (8) is trivial and it gives directly Result 3. In case (2) if one plugs $p_{S,N}(\boldsymbol{\lambda}^S) = \prod_{i \leqslant N} \rho_S(\lambda_i^S)$ in formula (11) the term $N^{-2} \ln p_{S,N}(\boldsymbol{\lambda}^s)$ is $o_N(1)$, so one may think that the prior has no influence on the formula. But this is not true, because the influence of this prior manifests itself through the data $\boldsymbol{Y}$ in the spherical integral which strongly depends on it. Going again through the four points above leading to the simplified Result 3, one can see that they all remain valid. And because the moments $\mathbb{E}(\lambda_1^S)^k$ are the same in scenarios (1) and (2) [and (3) as well], the last point based on the Nishimori identity (23) identifies the same maximizing $(m_{k,N}^*)$, i.e., the same optimal density $\rho^* = \rho_S$.

Let us provide an alternative information-theoretic counting argument in order to obtain Result 3 "directly" without going through all the previous steps, and that justifies *a posteriori* the equivalence of these seemingly very different situations at the level of the mutual information, which is thus insensitive to possible strong correlations between the eigenvalues of $\boldsymbol{S}$ and only depends on their density. By the chain rule for mutual information it can be decomposed as (recall $\boldsymbol{S} = \tilde{\boldsymbol{U}}^\dagger \boldsymbol{\lambda}^S \tilde{\boldsymbol{U}}$)

$$
\begin{aligned}
\frac{1}{N^2} I(\boldsymbol{Y}; \boldsymbol{S}) &= \frac{1}{N^2} I[\boldsymbol{Y}; (\tilde{\boldsymbol{U}}, \boldsymbol{\lambda}^S)] \\
&= \frac{1}{N^2} I(\boldsymbol{Y}; \tilde{\boldsymbol{U}} | \boldsymbol{\lambda}^S) + \frac{1}{N^2} I(\boldsymbol{Y}; \boldsymbol{\lambda}^S).
\end{aligned}
$$

Now, because there are only N eigenvalues while there are N^2 [resp. $N(N-1)/2$] angles defining the eigenbasis $\tilde{\boldsymbol{U}} \in \mathcal{U}(N)$ [resp. $\tilde{\boldsymbol{U}} \in \mathcal{O}(N)$], the second term in the right-hand side in the above decomposition is $O(1/N)$. Thus, at leading order $\Theta(N^2)$, the mutual information $I(\boldsymbol{Y}; \boldsymbol{S})$ and the one given the eigenvalues $I(\boldsymbol{Y}; \tilde{\boldsymbol{U}} | \boldsymbol{\lambda}^S)$ are equal. Said differently, there are so much fewer eigenvalues $\boldsymbol{\lambda}^S$ than angular degrees of freedom and data points that their inference has comparably negligible cost. In particular in $I(\boldsymbol{Y}; \tilde{\boldsymbol{U}} | \boldsymbol{\lambda}^S)$ the set of eigenvalues is given, so that their correlations does not matter and the mutual information can only depend on their density ρ_S. Since the priors (1)–(3) above have the same density the corresponding mutual informations are identical. Finally, note that by the same arguments we also have that $I(\boldsymbol{Y}; \tilde{\boldsymbol{U}} | \boldsymbol{\lambda}^S)$ [and thus $I(\boldsymbol{Y}; \boldsymbol{S})$ too]

is equal, at leading order in N, to

$$
I(\boldsymbol{Y}; \tilde{\boldsymbol{U}}) = I(\sqrt{\lambda} \, \tilde{\boldsymbol{U}}^\dagger \boldsymbol{\lambda}^S \tilde{\boldsymbol{U}} + \boldsymbol{\xi}; \tilde{\boldsymbol{U}}).
$$

E. A sanity check: The case of a Wigner signal

Consider the problem of denoising a Wigner matrix: $\boldsymbol{S}$ is itself a standard Wigner with same distribution as the noise $\boldsymbol{\xi}$. So $V(\boldsymbol{S}) = \boldsymbol{S}^2$ in (4). The data $\boldsymbol{Y}$ are therefore also a centered Wigner matrix with law

$$
P(\boldsymbol{Y}) \propto \exp \mathrm{Tr}\left[-\frac{\beta N}{4(1+\lambda)} \boldsymbol{Y}^2 \right]
$$

whose asymptotic spectral density is a semicircle of width $\sigma_Y := \sqrt{1+\lambda}$. This case is completely decoupled in the sense that each i.i.d. entry of the matrix $\boldsymbol{S}$ is corrupted independently by an i.i.d. Gaussian noise, so we should recover the known formulas for scalar decoupled Gaussian channels [91]. This can be verified as follows: in this case the supremum over ρ_s in Result 1 is attained for ρ_s being itself a semicircle of width $\sigma_s = 1$. Note that in this particular case, this can be deduced without making use of the Nishimori identity by realizing that whenever $\lambda \to +\infty$ or $\lambda \to 0_+$ it has to be so. Indeed in the noiseless limit $\lambda \to +\infty$ the posterior is peaked on the ground-truth signal $\boldsymbol{S}$ and thus a sample s will match it and have the same spectrum $\boldsymbol{\lambda}^s = \boldsymbol{\lambda}^S$ whose density is a semicircle of width 1. In the opposite completely noisy limit limit $\lambda = 0$, a sample from the posterior is simply drawn according to the prior $P_{S,N}$ which is the law of a standard Wigner matrix. Therefore in both cases the density ρ_s is a semicircle of width 1, but only in the second case the actual eigenvalues will match those of $\boldsymbol{S}$. For any intermediate value of λ the eigenvalues $\boldsymbol{\lambda}^s$ will be in a mixture that polarize more towards $\boldsymbol{\lambda}^S$ as λ increases, but which maintains the same asymptotic *density*. In the complex case $\beta = 2$, the asymptotic spherical integral $I^{(2)}[\rho_s, \rho_Y, \sqrt{\lambda}]$ has a closed expression when evaluated for two semicircle laws [102]:

$$
\begin{aligned}
I^{(2)}[\rho_s, \rho_Y, \sqrt{\lambda}] = \tfrac{1}{2}\{ &\sqrt{4\sigma(\lambda)^4 + 1} - 1 \\
&- \ln[1 + \sqrt{4\sigma(\lambda)^4 + 1}] + \ln 2 \}, \quad (26)
\end{aligned}
$$

where $\sigma(\lambda)^2 := \sqrt{\lambda} \sigma_Y \sigma_s = \sqrt{\lambda(1+\lambda)}$. Moreover, according to "Zuber's $\tfrac{1}{2}$-rule" [103] we can simply relate the real case $\beta = 1$ to the complex one $\beta = 2$:

$$
I^{(1)}[\rho_s, \rho_Y, \sqrt{\lambda}] = \tfrac{1}{2} I^{(2)}[\rho_s, \rho_Y, \sqrt{\lambda}]. \quad (27)
$$

Using that the second moment of the semicircle law $\mathbb{E}(\lambda_1^S)^2 = \int_{-2}^{2} dx \, x^2 \sqrt{4 - x^2}/(2\pi) = 1$ we reach from Result 3 the expected expression:

$$
\begin{aligned}
\frac{1}{N^2} I(\boldsymbol{Y}; \boldsymbol{S}) \to \frac{\beta}{4}\{ &2\lambda + 1 - \sqrt{4\lambda(1+\lambda) + 1} \\
&+ \ln[1 + \sqrt{4\lambda(1+\lambda) + 1}] - \ln 2 \} \\
= &\frac{\beta}{4} \ln(1+\lambda). \quad (28)
\end{aligned}
$$

The minimum mean-square error is thus

$$
\frac{1}{N^2} \mathbb{E} \|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2 \to \frac{1}{1+\lambda}. \quad (29)
$$

So we recover the formulas of [91]. Note that in the present case, the convergence $\to$ in the above identities are actually equalities for any N (but our derivation here is asymptotic in nature).

F. An explicit model with uniform spectral distribution

We consider model (2) with λ^S being a uniform permutation of equally spaced eigenvalues in $[-\sqrt{3}, \sqrt{3}]$:

$$p_{S,N}(\lambda^S) = \frac{1}{N!}\mathbf{1}\left\{\lambda^S \in \Pi\left[\sqrt{\gamma}\left(-\frac{1}{2}, \frac{1}{N}-\frac{1}{2}, \frac{2}{N}-\frac{1}{2}, \ldots, \frac{1}{2}-\frac{1}{N}\right)\right]\right\}, \tag{30}$$

where $\Pi(v)$ is the set of all $N!$ permutations of $v \in \mathbb{R}^N$, $\mathbf{1}(\cdot)$ is the indicator function, and $\gamma = \gamma_N \to 12$ enforces $\mathrm{Tr}\lambda_S^2 = N$. The advantage of this model is that the HCIZ integral appearing in Result 3 is explicit when $\beta = 2$. Let $\lambda^s \sim p_{S,N}$. The HCIZ integral (see Appendix A) does not depend on the ordering of the eigenvalues, therefore we can consider the increasing ordering $\lambda_i^s = \sqrt{\gamma}(i-1)/N - \sqrt{\gamma}/2$. Denote $\sigma := \gamma\lambda$. The HCIZ formula then gives (because the ratio of determinants is nonnegative we can insert an absolute value)

$$N^2 I_N^{(2)}(\lambda^s, \sqrt{\lambda}\lambda^Y, 1) = \ln\frac{\prod_{k\leqslant N-1}k!}{N^{N(N-1)/2}} + \ln\left|\frac{\det\left[\left(\exp\sqrt{\sigma}\lambda_j^Y\right)^{i-1}\exp\left(-N\sqrt{\sigma}\lambda_j^Y/2\right)\right]}{\Delta_N(\lambda^S)\Delta_N(\sqrt{\lambda}\lambda^Y)}\right|$$

$$= \ln\frac{\prod_{k\leqslant N-1}k!}{N^{N(N-1)/2}} + \ln\left|\frac{\det\left[\left(\exp\sqrt{\sigma}\lambda_j^Y\right)^{i-1}\right]}{\Delta_N(\lambda^S)\Delta_N(\sqrt{\lambda}\lambda^Y)}\right| - \frac{N}{2}\sqrt{\sigma}\,\mathrm{Tr}\lambda^Y.$$

The matrix $[(\exp\sqrt{\sigma}\lambda_j^Y)^{i-1}]$ is a generalized Vandermonde, and thus

$$\det\left[\left(\exp\sqrt{\sigma}\lambda_j^Y\right)^{i-1}\right] = \prod_{i<j}^{1,N}\left(\exp\sqrt{\sigma}\lambda_i^Y - \exp\sqrt{\sigma}\lambda_j^Y\right).$$

The mutual information from Result 3 then reads

$$\frac{1}{N^2}I(Y;S) = \lambda - \frac{1}{N^2}\sum_{i<j}^{1,N}\ln\left|\exp\sqrt{\sigma}\lambda_i^Y - \exp\sqrt{\sigma}\lambda_j^Y\right| + \frac{1}{N^2}\sum_{i<j}^{1,N}\ln\left(\sqrt{\lambda}|\lambda_i^Y - \lambda_j^Y|\right)$$

$$+ \frac{1}{N^2}\sum_{i<j}^{1,N}\ln\left(\sqrt{\gamma}\frac{|i-j|}{N}\right) + \frac{\sqrt{\sigma}}{2N}\mathrm{Tr}\lambda^Y - \frac{1}{N^2}\ln\frac{\prod_{k\leqslant N-1}k!}{N^{N(N-1)/2}} + o_N(1). \tag{31}$$

The MMSE can then be obtained using the I-MMSE relation (12):

$$\frac{\beta}{N^2}\mathbb{E}\|S - \mathbb{E}[S|Y]\|^2 = 4 - \frac{4}{N^2}\sum_{i<j}^{1,N}\frac{e^{\sqrt{\sigma}\lambda_i^Y}\frac{d}{d\lambda}\left(\sqrt{\sigma}\lambda_i^Y\right) - e^{\sqrt{\sigma}\lambda_j^Y}\frac{d}{d\lambda}\left(\sqrt{\sigma}\lambda_j^Y\right)}{e^{\sqrt{\sigma}\lambda_i^Y} - e^{\sqrt{\sigma}\lambda_j^Y}}$$

$$+ \frac{1}{\lambda} + \frac{4}{N^2}\sum_{i<j}^{1,N}\frac{\frac{d}{d\lambda}\left(\lambda_i^Y - \lambda_j^Y\right)}{\lambda_i^Y - \lambda_j^Y} + \sqrt{\frac{\gamma}{\lambda}}\frac{1}{N}\mathrm{Tr}\lambda^Y + \frac{2\sqrt{\sigma}}{N}\sum_{i\leqslant N}\frac{d}{d\lambda}\lambda_i^Y + o_N(1).$$

Introducing the Y-eigenvectors $Y\psi_i^Y = \lambda_i^Y\psi_i^Y$, the Hellmann-Feynman theorem implies

$$\frac{d}{d\lambda}\lambda_i^Y = \frac{1}{2\sqrt{\lambda}}\left(\psi_i^Y\right)^\dagger S\psi_i^Y =: \frac{1}{2\sqrt{\lambda}}p_i,$$

where S is the ground truth in (2) (not to be confused with s, another independent sample from $P_{S,N}$). As a consequence we finally obtain the explicit expression

$$\frac{\beta}{N^2}\mathbb{E}\|S - \mathbb{E}[S|Y]\|^2 = 4 - \frac{2\sqrt{\gamma}}{N^2}\sum_{i<j}^{1,N}\frac{e^{\sqrt{\sigma}\lambda_i^Y}\left(\frac{1}{\sqrt{\lambda}}\lambda_i^Y + p_i\right) - e^{\sqrt{\sigma}\lambda_j^Y}\left(\frac{1}{\sqrt{\lambda}}\lambda_j^Y + p_j\right)}{e^{\sqrt{\sigma}\lambda_i^Y} - e^{\sqrt{\sigma}\lambda_j^Y}}$$

$$+ \frac{1}{\lambda} + \frac{2}{\sqrt{\lambda}N^2}\sum_{i<j}^{1,N}\frac{p_i - p_j}{\lambda_i^Y - \lambda_j^Y} + \sqrt{\frac{\gamma}{\lambda}}\frac{1}{N}\mathrm{Tr}\lambda^Y + \frac{\sqrt{\gamma}}{N}\sum_{i\leqslant N}p_i + o_N(1). \quad (32)$$

Let us introduce the asymptotic spectral densities ρ_s and ρ_Y associated with the matrices s and Y. Then the above expression reads, in the large size limit $N \to +\infty$,

$$\frac{1}{N^2}I(Y;S) \to \lambda + \frac{\ln\lambda\gamma}{4} + \frac{1}{2}\int d\rho_Y(x)\,d\rho_Y(y)$$

$$\times \ln\left|\frac{x - y}{\exp x\sqrt{\lambda\gamma} - \exp y\sqrt{\lambda\gamma}}\right|. \quad (33)$$

We used that

$$\frac{1}{N^2}\sum_{i<j}^{1,N}\ln\frac{|i-j|}{N} \to \frac{1}{2}\int_{[0,1]\times[0,1]}dx\,dy\ln|x-y|$$

$$= -\frac{3}{4} = \lim_{N\to+\infty}\frac{1}{N^2}\ln\frac{\prod_{k\leqslant N-1}k!}{N^{N(N-1)/2}}$$

so these two terms asymptotically cancel each others. We also used that Y, as a sum of asymptotically traceless matrices, is asymptotically traceless too and therefore the term $\int d\rho_Y(x)x = 0$.

Note that, as explained below Result 3, we could have fixed from the beginning one arbitrary permutation of the eigenvalues: $p_{S,N}(\lambda^S) = \delta[\lambda^S - \sqrt{\gamma}(-1/2, 1/N - 1/2, 2/N - 1/2, \ldots, 1/2 - 1/N)]$, instead of considering the uniform measure (30) over permutations. This would have lead to the same calculations as can be easily seen. What is less trivial to see (because in that case we cannot simplify anymore the HCIZ formula using the generalized Vandermonde form) is that the result would be asymptotically the same if the prior was instead uniform but not necessarily equally spaced, i.e., $p_{S,N} = \mathcal{U}[-\sqrt{3}, \sqrt{3}]^{\otimes N}$.

The λ-dependent data spectral distribution ρ_Y can be obtained from free probability as follows; we refer to [89,104] for clean definitions, domains of definitions and properties of the functions we are going to use now. The complex-valued Green function (or minus Stieljes transform) associated with ρ, whose domain is the complex plane minus the support of ρ, is

$$G_\rho(z) := \int d\rho(x)\frac{1}{z - x}.$$

The Blue function is its functional inverse verifying $B_\rho[G_\rho(z)] = G_\rho[B_\rho(z)] = z$. Then the complex valued R transform is defined as

$$R_\rho(z) = B_\rho(z) - \frac{1}{z} = \sum_{i\geqslant 1}k_i z^{i-1},$$

where the coefficients $(k_i)_{i\geqslant 1}$ in its series expansion are the so-called free cumulants associated with density ρ. Asymptotically, the matrix $\sqrt{\lambda}S$ has eigenvalue density $\rho_{\sqrt{\lambda}S}$ which is the uniform distribution in $[-\sqrt{3\lambda}, \sqrt{3\lambda})$. The associated Green function is

$$G_{\rho_{\sqrt{\lambda}S}}(z) = \frac{1}{2\sqrt{3\lambda}}\ln\frac{z + \sqrt{3\lambda}}{z - \sqrt{3\lambda}},$$

thus

$$\mathcal{R}_{\rho_{\sqrt{\lambda}S}}(z) = \sqrt{3\lambda}\coth(z\sqrt{3\lambda}) - \frac{1}{z}.$$

The R transform of the standard Wigner semicircle law is the identity: $R_{\rho_Z}(z) = z$. Finally, by additivity of the R transform for asymptotically free random matrices, the R transform of the spectral density of the data matrix is

$$R_{\rho_Y}(z) = R_{\rho_{\sqrt{\lambda}S}}(z) + R_{\rho_Z}(z) = \sqrt{3\lambda}\coth(z\sqrt{3\lambda}) - \frac{1}{z} + z.$$

Its Blue function is thus $B_{\rho_Y}(z) = \sqrt{3\lambda}\coth(z\sqrt{3\lambda}) + z$ from which we get a transcendental equation for its Green function:

$$z = \sqrt{3\lambda}\coth(G_{\rho_Y}(z)\sqrt{3\lambda}) + G_{\rho_Y}(z). \quad (34)$$

This equation can be solved numerically using a complex nonlinear solver. A MATHEMATICA code to do so is provided in Appendix C. From its solution we can access the spectral density due to

$$\rho_Y(x) = \frac{1}{\pi}\lim_{\varepsilon\to 0_+}|\mathrm{Im}G_{\rho_Y}(x - i\varepsilon)|. \quad (35)$$

Figure 1 shows in red the asymptotic prediction from the spectrum extracted from the numerical solution of (34) and (35). It almsot perfectly matches the empirical density of eigenvalues of Y for realizations of the model for large sizes, see the blue histograms.

Given a signal-to-noise ratio λ, we can compute the mutual information as $N \to +\infty$ using (33), (34) and (35). This yields the pink dots of Fig. 2. The blue dots are instead for the mutual information computed for large realizations of the model (2) for a S with a uniform spectrum; see formula (31). The orange dots are for the MMSE for that case; see formula (32). These curves are compared to the case of Wigner signal and match surprisingly well up to relative differences of $O(10^{-3})$. But our perturbative expansions of the next section as well as the comparison with the asymptotic predictions from (33) show that this difference, even if small, is not just due to numerical imprecisions: the curves really are different even in the large size limit. Yet it is interesting to observe
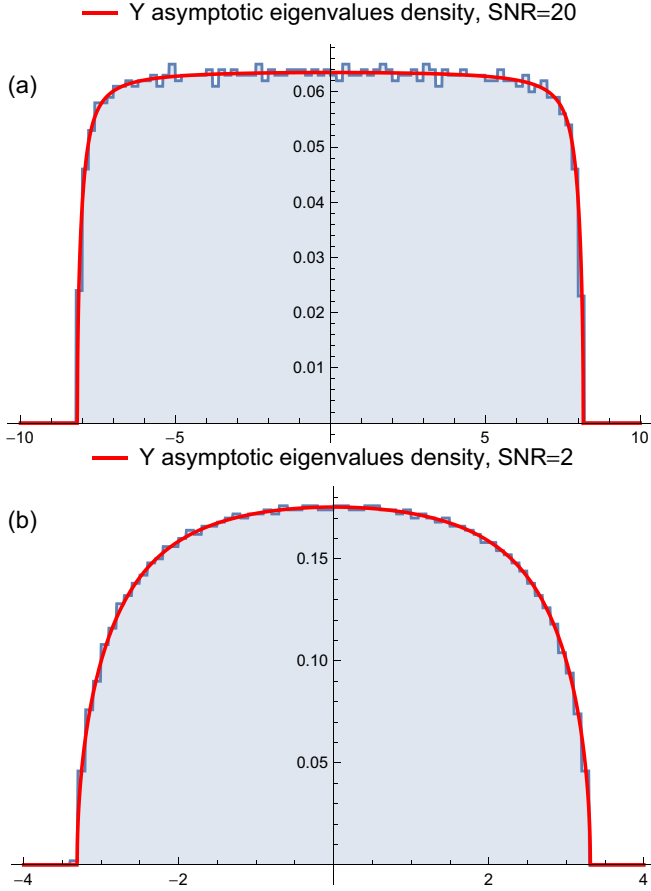
FIG. 1. (a) Asymptotic $N \to +\infty$ spectral density $\rho_Y(x)$ (red) for the denoising model (2) with $\lambda = 20$ and a signal $\boldsymbol{S}$ with uniform eigenvalues in $[-\sqrt{3}, \sqrt{3})$. It is compared to the empirical spectral density of $\boldsymbol{Y}$ for a realization of size $N = 5000$ (blue). (b) The same for a smaller signal-to-noise ratio $\lambda = 2$. As expected, the spectrum resembles more the semicircle law in that case. The density does approach Wigner's semicircle law of radius 2 as $\lambda \to 0_+$.

that the simple (decoupled) case of matrix denoising with $\boldsymbol{S}$ a Wigner matrix allows us to very accurately approximate the information-theoretic quantities of the much less trivial setting where $\boldsymbol{S}$ has a uniform spectrum (and therefore the matrix elements of $\boldsymbol{S}$ are dependent, as opposed to the Wigner case). Further investigation around this fact is needed and left for future work.

III. PERTURBATIVE EXPANSIONS FOR HERMITIAN MATRIX DENOISING

We exploit perturbative expansions of the HCIZ integral for small λ to discuss the corresponding expansions of the mutual information and MMSE as predicted by Result 3. In this section we consider only complex cases $\beta = 2$, for which expansions of the HCIZ integral have been worked out [105,106].

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ two Hermitian $N \times N$ matrices. We use an expansion of the HCIZ integral (9) in the complex case
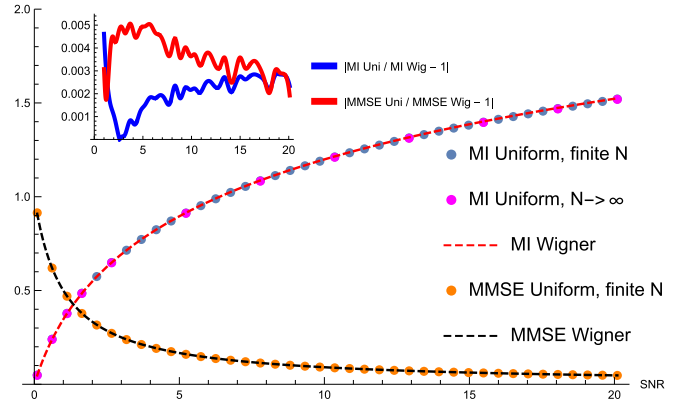

FIG. 2. Main panel: The abcissa corresponds to the signal-to-noise ratio λ in model (2). The blue dots correspond to the mutual information (MI) for the uniform spectrum case evaluated from (31) for $N = 1000$ averaged over 100 independent realizations; the orange dots are the MMSE in the same Monte Carlo experiment, evaluated from (32). The pink dots correspond to the asymptotic $N \to +\infty$ mutual information for the uniform spectrum case evaluated from (33). The finite N and asymptotic $N \to +\infty$ values of the mutual information match very closely as can be seen from the superposition of the pink and blue dots. The red dashed line is the mutual information for the Wigner signal case (28) and the black one the MMSE (29). All is for $\beta = 2$. The curves for the uniform and semicircle laws match surprisingly well but are actually different. Inset: These curves quantify the relative difference between the empirical curves (the blue or orange dots) for the uniform case and the (dashed) curves for the Wigner case. The relative difference is typically of order $O(10^{-3})$. When comparing instead the $N \to +\infty$ curve for the uniform case (pink dots) to the Wigner mutual information so that any finite-size effects are removed, a difference of the same order survives (which is much higher than the expected numerical precision for these computations). This confirms that the curves are *not* exactly the same.

$I_N^{(2)}(\boldsymbol{A}, \boldsymbol{B}, \sqrt{\lambda})$ in terms of moments

$$\theta_p := \lim_{N \to +\infty} \frac{1}{N} \mathrm{Tr} \boldsymbol{A}^p = \lim_{N \to +\infty} \frac{1}{N} \sum_{i \leqslant N} \left(\lambda_i^A\right)^p,$$

$$\bar{\theta}_p := \lim_{N \to +\infty} \frac{1}{N} \mathrm{Tr} \boldsymbol{B}^p = \lim_{N \to +\infty} \frac{1}{N} \sum_{i \leqslant N} \left(\lambda_i^B\right)^p$$

for integer $p \geqslant 1$. Note that by concentration θ_p is also equal to $\lim_{N \to +\infty} N^{-1} \mathbb{E} \mathrm{Tr} \boldsymbol{A}^p$ and similarly for $\bar{\theta}_p$. We assume that $\boldsymbol{A}$ and $\boldsymbol{B}$ are *traceless*, i.e., $\theta_1 = \bar{\theta}_1 = 0$. Then, according to [106],

$$\lim_{N \to +\infty} I_N^{(2)}(\boldsymbol{A}, \boldsymbol{B}, \sqrt{\lambda}) = I^{(2)}[\rho_A, \rho_B, \sqrt{\lambda}]$$

$$= \sum_{n \geqslant 2} \lambda^{\frac{n}{2}} F_n(\boldsymbol{A}, \boldsymbol{B}) \qquad (36)$$

with (terms up to $n = 8$ are explicitly derived in [106] and diagrammatic rules are given for higher orders; see also Appendix C for their complete expressions)

$$F_2 = \tfrac{1}{2}\theta_2\bar{\theta}_2,$$

$$F_3 = \tfrac{1}{3}\theta_3\bar{\theta}_3, \qquad (37)$$

$$F_4 = \tfrac{3}{4}\theta_2\bar{\theta}_2 - \tfrac{1}{2}\theta_2^2\bar{\theta}_4 - \tfrac{1}{2}\theta_4\bar{\theta}_2^2 + \tfrac{1}{4}\theta_4\bar{\theta}_4.$$

We will also make use of the derivatives with respect to the moments $\bar{\theta}_p$. These read

$$\frac{\partial}{\partial \bar{\theta}_p} I^{(2)}[\rho_A, \rho_B, \sqrt{\lambda}] = \frac{\bar{D}_p}{p} \quad \text{for} \quad p \geqslant 2,$$

with

$$\begin{aligned}
\bar{D}_2 &= \lambda \theta_2 + \lambda^2 \tfrac{1}{2}(3\theta_2 - 4\theta_4 \bar{\theta}_2) + O(\lambda^{5/2}), \\
\bar{D}_3 &= \lambda^{3/2} \theta_3 + O(\lambda^{5/2}), \\
\bar{D}_4 &= -\lambda^2 (2\theta_2^2 - \theta_4) + O(\lambda^3),
\end{aligned} \tag{38}$$

where the higher order corrections come from the structure of F_5 and F_6 and can be worked out from [106]. These formulas are applied for $A = S$ and $B = Y = \sqrt{\lambda}S + \xi$ with $N^{-1}\mathrm{Tr}S = o_N(1)$ and $N^{-1}\mathrm{Tr}Y = o_N(1)$. As explained in Remark 1 after Result 3 the mutual information remains the same if we center the signal to make S and Y traceless. Although this is not necessary, and one can work out the expansion for a noncentered signal and data, this turns out to be a major simplification in the ensuing calculations. For the mutual information, according to Result 3 we find the expansion when $\beta = 2$:

$$\lim_{N \to +\infty} \frac{1}{N^2} I(Y; S) = \lambda \theta_2 - \sum_{n \geqslant 2} \lambda^{\frac{n}{2}} F_n \tag{39}$$

and for the MMSE

$$\begin{aligned}
\lim_{N \to +\infty} &\frac{\beta}{N^2} \mathbb{E} \|S - \mathbb{E}[S|Y]\|^2 \\
&= 4\theta_2 - 4 \frac{\partial}{\partial \lambda} I^{(2)}[\rho_S, \rho_Y, \sqrt{\lambda}] \\
&\quad - 4 \sum_{p \geqslant 2} \frac{\partial}{\partial \bar{\theta}_p} I^{(2)}[\rho_S, \rho_Y, \sqrt{\lambda}] \frac{d\bar{\theta}_p}{d\lambda} \\
&= 4\theta_2 - 4 \sum_{n \geqslant 2} \frac{n}{2} \lambda^{\frac{n}{2}-1} F_n - 4 \sum_{p \geqslant 2} \frac{\bar{D}_p}{p} \frac{d\bar{\theta}_p}{d\lambda}. \tag{40}
\end{aligned}$$

In these expressions F_n and D_p are given by their expansions in terms of the moments

$$\begin{aligned}
\theta_p &= \lim_{N \to +\infty} N^{-1}\mathrm{Tr}S^p = \lim_{N \to +\infty} N^{-1}\mathbb{E}\mathrm{Tr}S^p, \\
\bar{\theta}_p &= \lim_{N \to +\infty} N^{-1}\mathrm{Tr}Y^p = \lim_{N \to +\infty} N^{-1}\mathbb{E}\mathrm{Tr}Y^p,
\end{aligned}$$

which themselves are polynomials in $\sqrt{\lambda}$. To go further we must fix a specific model of interest.

Example 1 (Wigner signal): Let $\xi = \xi^\dagger \in \mathbb{C}^{N \times N}$, $\xi \sim \exp \mathrm{Tr}[-\frac{N}{2}\xi^2]$ a standard Hermitian Wigner matrix; this corresponds to a potential $V(x) = x^2$ in (4). Take an i.i.d. copy ξ' and set $S = \xi'$ and $Y = \sqrt{\lambda}\xi' + \xi$. We note that $Y \sim \sqrt{1+\lambda}\xi$. Wigner's semicircle law

$$\rho_\xi(x) = \mathbf{1}(|x| \leqslant 2) \frac{\sqrt{4-x^2}}{2\pi}$$

implies for even moments (odd moments vanish)

$$\begin{aligned}
\theta_{2p} &= \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}\mathrm{Tr}\xi^{2p} = \frac{1}{p+1}\binom{2p}{p}, \\
\bar{\theta}_{2p} &= \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}\mathrm{Tr}Y^{2p} = \frac{(1+\lambda)^p}{p+1}\binom{2p}{p}.
\end{aligned}$$

From $\theta_2 = 1, \theta_4 = 2, \theta_6 = 5$ and $\bar{\theta}_2 = 1 + \lambda, \bar{\theta}_4 = 2(1 + \lambda)^2, \bar{\theta}_6 = 5(1 + \lambda)^3$ we find

$$F_2 = \tfrac{1}{2}(1+\lambda), \quad F_4 = -\tfrac{1}{4}(1+\lambda)^2, \quad \text{and} \quad F_6 = \tfrac{1}{3}(1+\lambda)^3.$$

The expansion (39) for the mutual information yields

$$\lim_{N \to +\infty} \frac{1}{N^2} I(Y; S) = \frac{\lambda}{2} - \frac{\lambda^2}{4} + \frac{\lambda^3}{6} + O(\lambda^4). \tag{41}$$

We recognize the expansion of $\frac{1}{2}\ln(1+\lambda)$ and the result is consistent with (28).

Example 2 (Wishart signal): Consider $S = XX^\dagger$ and $Y = \sqrt{\lambda}XX^\dagger + \xi$, where the noise $\xi \in \mathbb{C}^{N \times N}$ is an Hermitian Wigner matrix normalized as in the previous example and $X \in \mathbb{C}^{N \times M}$ is drawn from $P_{X,M}(X) \propto \exp \mathrm{Tr}[-MXX^\dagger]$. Let $\varphi := N/M$. For $\varphi \leqslant 1$ the eigenvalue j.p.d.f. is well defined, and corresponds to the potential $V(x) = 2(1 - 1/\varphi)\ln|x| + 2x/\varphi$ in (4) (see, e.g., [104]). When $\varphi > 1$ the matrix is rank deficient and there is no well-defined j.p.d.f. for the eigenvalues but the model can be regularized as explained in Sec. II and the final conjectures apply. In particular the Marcenko-Pastur distribution for the eigenvalues of $S = XX^\dagger$ is well defined for all $\varphi > 0$:

$$\begin{aligned}
\rho_{\mathrm{MP}}(x) = &\max(1 - 1/\varphi, 0)\delta(x) \\
&+ \frac{\mathbf{1}(c \leqslant x \leqslant d)}{2\pi \varphi x}\sqrt{(x-c)(d-x)}, \tag{42}
\end{aligned}$$

where $c := (\sqrt{\varphi} - 1)^2$ and $d := (\sqrt{\varphi} + 1)^2$. The spectral moments are deduced by standard integration methods, and we find

$$\begin{aligned}
\lim_{N \to +\infty} &\frac{1}{N} \mathrm{Tr}(XX^\dagger)^p \\
&= \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}\mathrm{Tr}(XX^\dagger)^p = \frac{1}{p} \sum_{k \leqslant p} \varphi^{k-1} \binom{p}{k}\binom{p}{k-1}.
\end{aligned}$$

In particular $\lim_{N \to +\infty} N^{-1}\mathbb{E}\mathrm{Tr}XX^\dagger = 1$ for the first moment. Now, as explained before, in order to compute the mutual information it is convenient to center the signal so that it becomes traceless. In other words we replace $XX^\dagger$ by $S = XX^\dagger - \mathrm{I}_{d,N}$ so that $N^{-1}\mathrm{Tr}S \to 0$. This also implies $Y = \sqrt{\lambda}(XX^\dagger - \mathrm{I}_{d,N}) + \xi$ and $N^{-1}\mathrm{Tr}Y \to 0$. The first moments of the spectral density of this S are in the asymptotic limit $\theta_2 = \varphi, \theta_3 = \varphi^2, \theta_4 = \varphi^3 + 2\varphi^2, \theta_5 = \varphi^4 + 5\varphi^3, \theta_6 = 5\varphi^3 + 9\varphi^4 + \varphi^5$ and those of Y are $\bar{\theta}_1 = 0, \bar{\theta}_2 = 1 + \lambda\varphi, \bar{\theta}_3 = \lambda^{3/2}\varphi^2, \bar{\theta}_4 = \lambda^2(\varphi^3 + 2\varphi^2) + 4\lambda\varphi + 2, \bar{\theta}_5 = O(\lambda^{3/2}), \bar{\theta}_6 = 5 + O(\lambda)$. This yields

$$\begin{aligned}
F_2 &= \frac{\varphi}{2} + \lambda\frac{\varphi^2}{2}, \\
F_3 &= \lambda^{3/2}\frac{\varphi^4}{3}, \\
F_4 &= -\frac{\varphi^4}{4} - \lambda\frac{\varphi^3}{2} + \lambda^2\frac{1}{4}(\varphi^6 - \varphi^4), \tag{43} \\
F_5 &= O(\lambda^{3/2}), \\
F_6 &= \frac{\varphi^3}{3} - \frac{\varphi^4}{6} + O(\lambda).
\end{aligned}$$

For the mutual information we find

$$\lim_{N\to+\infty}\frac{1}{N^2}I(\mathbf{Y};\mathbf{S})$$

$$= \lambda\frac{\varphi}{2} - \lambda^2\frac{\varphi^2}{4} + \lambda^3\frac{\varphi^3 - \varphi^4}{6} + O(\lambda^4). \qquad (44)$$

We note that the contribution of the order $O(\lambda^3)$ only comes from the order $O(\lambda)$ in F_4 and the constant term in F_6. We also remark that for $\varphi = 1$ the first two orders are the same than the pure Wigner case of Example 1. It is possible to show that this is a universal feature for all matrices $\mathbf{S}$ such that $N^{-1}\mathrm{Tr}\mathbf{S} \to 0$ and $N^{-1}\mathrm{Tr}\mathbf{S}^2 \to 1$; see the next example.

Example 3 (general case): As before $(\theta_p)_{p\geqslant 1}$ correspond to the asymptotic spectral moments of the signal $\mathbf{S}$. It easy to skim through the above calculations and obtain the first terms of an expansion for general signals with even spectral density such that $\theta_2 = 1$ and $\theta_{2p+1} = 0$, $p \geqslant 0$ (note that for the traceless Marchenko-Pastur distribution this is true only if $p = 0$). At third order the resulting expansion can be read off by removing the contribution of F_3 and the term $-\varphi^4/6$ from F_6, and we find

$$\lim_{N\to+\infty}\frac{1}{N^2}I(\mathbf{Y};\mathbf{S}) = \frac{\lambda}{2} - \frac{\lambda^2}{4} + \frac{\lambda^3}{6} + O(\lambda^4). \qquad (45)$$

The generic case until fourth order is heavy to handle by hand. We provide in Appendix C a MATHEMATICA code to get the following expansions. Let $(k_p)_{p\geqslant 2}$ be the free cumulants associated with the asymptotic spectral density ρ_S of $\mathbf{S}$ (see, e.g., [4,89] to learn about free cumulants). For $N \to +\infty$ followed by $\lambda \to 0_+$ and a ρ_S such that $\theta_1 = k_1 = 0$ and $\theta_2 = k_2 = 1$,

$$\lim_{N\to+\infty}\frac{1}{N^2}I(\mathbf{Y};\mathbf{S})$$

$$= \frac{\lambda}{2} - \frac{\lambda^2}{4} + \lambda^3\frac{1 - k_3^2}{6} - \lambda^4\frac{1 + 4k_3^2 + k_4^2}{8} + o(\lambda^4). \qquad (46)$$

Or expressed in terms of the moments (the mapping between moments and free cumulants can be obtained using the routines in Appendix C),

$$\lim_{N\to+\infty}\frac{1}{N^2}I(\mathbf{Y};\mathbf{S}) = \frac{\lambda}{2} - \frac{\lambda^2}{4} + \lambda^3\frac{1 - \theta_3^2}{6}$$

$$- \lambda^4\frac{5 + 4\theta_3^2 + \theta_4^2 - 4\theta_4}{8} + o(\lambda^4). \qquad (47)$$

Interestingly, for any even spectral density this matches the pure Wigner case of Example 1 up to third order. However, this breaks down at fourth order as soon as the fourth moment θ_4 of ρ_S is different from 2. This is another indication that the curves for the uniform spectrum and Wigner cases of Fig. 2 are different. Indeed, it can be checked that the

(F_n) in the expansion (39) for these two cases are very close but different—or that their respective expansions (47) are the same up to order three, but the order four for the Wigner case is $-\frac{1}{8}\lambda^4$ while it is $-\frac{1}{8}\frac{26}{25}\lambda^4$ for the case of uniform spectrum in $[-\sqrt{3}, \sqrt{3})$, and is thus very close. If one is instead interested in obtaining expansions in the opposite large signal-to-noise ratio regime, it is possible to do so using the expressions found in [102].

IV. DENOISING OF A ROTATIONALLY INVARIANT MATRIX: NON-HERMITIAN CASE

A. The model

We consider again a model of the form (2) but we now relax the hypothesis that $\mathbf{S}$ is Hermitian. This time we consider that $\boldsymbol{\xi}$ is a standard (non-Hermitian) Ginibre matrix with law

$$dP_{\xi,N}(\boldsymbol{\xi}) = C_N\, d\boldsymbol{\xi}\, \exp\mathrm{Tr}\left(-\frac{\beta N}{2}\boldsymbol{\xi}\boldsymbol{\xi}^\dagger\right).$$

Its entries are typically of order $O(1/\sqrt{N})$ and singular values $O(1)$. The planted full-rank matrix signal $\mathbf{S} \in \mathbb{K}^{N\times N}$ is no longer Hermitian but is still rotationally invariant in the sense that

$$dP_{S,N}(\mathbf{S}) = dP_{S,N}(\mathbf{O}\mathbf{S}\tilde{\mathbf{O}})$$

for any orthogonal or unitary $\mathbf{O}, \tilde{\mathbf{O}}$. It has $O(1/\sqrt{N})$ entries and $O(1)$ singular values. Its singular values decomposition (SVD) reads $\mathbf{S} = \tilde{\mathbf{U}}\boldsymbol{\sigma}^S\tilde{\mathbf{V}}$. Its singular values $\boldsymbol{\sigma}^S$ have a generic empirical distribution converging as $N \to +\infty$ to ρ_S with finite support. Left and right rotational invariance implies that its measure is decomposed as

$$dP_{S,N}(\mathbf{S}) \propto d\mu_N^{(\beta)}(\tilde{\mathbf{U}})\, d\mu_N^{(\beta)}(\tilde{\mathbf{V}})\, dp_{S,N}(\boldsymbol{\sigma}^S),$$

where the Vandermonde determinant and other terms inherent to the change of variable are included in the generic symmetric j.p.d.f. $p_{S,N}(\boldsymbol{\sigma}^S)$ of the singular values. For example, in the case of a measure defined by a rotationally invariant potential it reads [107]

$$dP_{S,N}(\mathbf{S}) \propto d\mu_N^{(\beta)}(\tilde{\mathbf{U}})\, d\mu_N^{(\beta)}(\tilde{\mathbf{V}})\, d\boldsymbol{\sigma}^S\left(\prod_{i\leqslant N}\sigma_i^S\right)^{\beta-1}$$

$$\times \exp\mathrm{Tr}\left[-\frac{\beta N}{2}V(\boldsymbol{\sigma}^S)\right]\left|\Delta_N(\boldsymbol{\sigma}_S^2)\right|^\beta. \qquad (48)$$

B. Free entropy through random matrix analysis

Using the SVD $\mathbf{s} = \mathbf{U}\boldsymbol{\sigma}^s\mathbf{V}$ the free entropy reads

$$f_N := \frac{1}{N^2}\ln\int dP_{S,N}(\mathbf{s})\exp\frac{\beta N}{2}\mathrm{Tr}(\sqrt{\lambda}\mathbf{Y}^\dagger\mathbf{s} + \sqrt{\lambda}\mathbf{Y}\mathbf{s}^\dagger - \lambda\mathbf{s}^\dagger\mathbf{s})$$

$$\propto \frac{1}{N^2}\ln\int dp_{S,N}(\boldsymbol{\sigma}^s)\exp\left[-\frac{\beta\lambda N}{2}\mathrm{Tr}\boldsymbol{\sigma}_s^2 + (\beta - 1)\sum_{i\leqslant N}\ln\sigma_i^s\right]$$

$$\times \int d\mu_N^{(\beta)}(\mathbf{U})\, d\mu_N^{(\beta)}(\mathbf{V})\exp\beta\sqrt{\lambda}N\mathrm{ReTr}(\mathbf{Y}^\dagger\mathbf{U}\boldsymbol{\sigma}^s\mathbf{V})$$

$$= \frac{1}{N^2}\ln\int d\boldsymbol{\sigma}^s\exp N^2\left[\frac{1}{N^2}\ln p_{S,N}(\boldsymbol{\sigma}^s) - \frac{\beta\lambda}{2N}\mathrm{Tr}\boldsymbol{\sigma}_s^2 + J_N^{(\beta)}(\boldsymbol{\sigma}^s, \boldsymbol{\sigma}^Y, 2\sqrt{\lambda})\right] + O(1/N).$$

Here and everywhere integrals over individual singular values are restricted to $\mathbb{R}_{\geqslant 0}$. The expression of the rectangular log-spherical integral density is [108,109]

$$
\begin{aligned}
J_N^{(\beta)}(\boldsymbol{A}, \boldsymbol{B}, \gamma) &= J_N^{(\beta)}(\boldsymbol{\sigma}^A, \boldsymbol{\sigma}^B, \gamma) \\
&:= \frac{1}{N^2} \ln \int d\mu_N^{(\beta)}(\boldsymbol{U}) \, d\mu_N^{(\beta)}(\boldsymbol{V}) \\
&\quad \times \exp \frac{\beta\gamma}{2} N \mathrm{Re}\mathrm{Tr}\,(\boldsymbol{\sigma}^A \boldsymbol{U} \boldsymbol{\sigma}^B \boldsymbol{V}) \qquad (49)
\end{aligned}
$$

for generic $N \times N$ matrices $\boldsymbol{A}$, $\boldsymbol{B}$ with respective singular values $\boldsymbol{\sigma}^A$ and $\boldsymbol{\sigma}^B$. It has a well-defined limit [110,111]:

$$
J^{(\beta)}[\rho_A, \rho_B, \gamma] := \lim_{N \to +\infty} J_N^{(\beta)}(\boldsymbol{\sigma}^A, \boldsymbol{\sigma}^B, \gamma), \qquad (50)
$$

where ρ_A, ρ_B are the asymptotic normalized densities of singular values associated with $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Let ρ_Y be the the asymptotic singular values density associated with the data $\boldsymbol{Y}$; again both $\boldsymbol{\sigma}^Y$ and ρ_Y are obtainable. In the large size limit $N \to +\infty$ we obtain by saddle-point the following conjecture for $f_N(\boldsymbol{Y})$:

$$
\begin{aligned}
f_N = \sup_{\boldsymbol{\sigma}^s \in \mathbb{R}_{\geqslant 0}^N} \Bigg[&\frac{1}{N^2} \ln p_{S,N}(\boldsymbol{\sigma}^s) - \frac{\beta\lambda}{2N} \mathrm{Tr}\boldsymbol{\sigma}_s^2 \\
&+ J_N^{(\beta)}(\boldsymbol{\sigma}^s, \boldsymbol{\sigma}^Y, 2\sqrt{\lambda}) \Bigg] + \tau_N.
\end{aligned}
$$

Focusing on the case of a prior of the form (48),

$$
\begin{aligned}
f_N = \sup_{\boldsymbol{\sigma}^s \in \mathbb{R}_{\geqslant 0}^N} \Bigg[&\frac{\beta}{2} \sum_{i \neq j}^{1,N} \frac{\ln \left| (\sigma_i^s)^2 - (\sigma_j^s)^2 \right|}{N^2} \\
&- \frac{\beta}{2} \sum_{i \leqslant N} \frac{\lambda (\sigma_i^s)^2 + V(\sigma_i^s)}{N} + J_N^{(\beta)}(\boldsymbol{\sigma}^s, \boldsymbol{\sigma}^Y, 2\sqrt{\lambda}) \Bigg] + \tau_N.
\end{aligned}
$$

Introducing asymptotic densities of singular values and by the same type of LDP as in Sec. II C, the previous variational formula can be reexpressed as

$$
\begin{aligned}
f_N \to \sup_{\rho_s \in \mathcal{P}_{\geqslant 0}} \Bigg\{ &\frac{\beta}{2} \int d\rho_s(x) \, d\rho_s(y) \ln |x^2 - y^2| - \frac{\beta}{2} \int d\rho_s(x) \\
&\times [\lambda x^2 + V(x)] + J^{(\beta)}[\rho_s, \rho_Y, 2\sqrt{\lambda}] \Bigg\} + \tau.
\end{aligned}
$$

The optimization is over a p.d.f. with bounded nonnegative support. The constants τ_N and τ are fixed by the constraint $f_N(\lambda = 0) = 0$:

$$
\begin{aligned}
\tau_N := - \sup_{\boldsymbol{\sigma}^s \in \mathbb{R}_{\geqslant 0}^N} \Bigg[&\frac{\beta}{2} \sum_{i \neq j}^{1,N} \frac{\ln \left| (\sigma_i^s)^2 - (\sigma_j^s)^2 \right|}{N^2} \\
&- \frac{\beta}{2} \sum_{i \leqslant N} \frac{V(\sigma_i^s)}{N} \Bigg] + o_N(1),
\end{aligned}
$$

$$
\begin{aligned}
\tau := - \sup_{\rho_s \in \mathcal{P}_{\geqslant 0}} \Bigg[&\frac{\beta}{2} \int d\rho_s(x) \, d\rho_s(y) \ln |x^2 - y^2| \\
&- \frac{\beta}{2} \int d\rho_s(x) V(x) \Bigg].
\end{aligned}
$$

Again, as f_N ends up being solely a function of the singular values of the data matrix, it is expected to be self-averaging with respect to $\boldsymbol{Y}$: $\mathbb{E} f_N = f_N + o_N(1)$. The free entropy is linked to the mutual information through

$$
I(\boldsymbol{Y}; \boldsymbol{S}) = -\mathbb{E} f_N + \frac{\beta\lambda N}{2} \mathbb{E}\mathrm{Tr}\boldsymbol{S}\boldsymbol{S}^\dagger.
$$

With the Ginibre noise instead of Wigner and for a non-Hermitian signal $\boldsymbol{S}$ the I-MMSE relation reads

$$
\frac{1}{N^2} \mathbb{E}\|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2 = \frac{2}{\beta N^2} \frac{d}{d\lambda} I(\boldsymbol{Y}; \boldsymbol{S}). \qquad (51)
$$

Like in the Hermitian case, the Nishimori identities combined with the concentration of the moments of the density of singular values of the posterior samples imply together that the supremum is attained for the density of singular values of the planted signal $\boldsymbol{S}$.

Result 4 (Mutual information of rotationally invariant matrix denoising). Let the singular values $\boldsymbol{\sigma}^s \sim p_{S,N}$ of a random matrix drawn according to the prior $P_{S,N}$. The mutual information of model (2) in the case where $\boldsymbol{S}$ is not necessarily Hermitian and the noise $\boldsymbol{\xi}$ is a standard Ginibre matrix verifies

$$
\frac{1}{N^2} I(\boldsymbol{Y}; \boldsymbol{S}) = \frac{\beta\lambda}{N} \mathrm{Tr}\boldsymbol{\sigma}_s^2 - J_N^{(\beta)}(\boldsymbol{\sigma}^s, \boldsymbol{\sigma}^Y, 2\sqrt{\lambda}) + o_N(1).
$$

Introducing asymptotic densities of singular values it reads as $N \to +\infty$

$$
\frac{1}{N^2} I(\boldsymbol{Y}; \boldsymbol{S}) \to \beta\lambda \int d\rho_s(x) \, x^2 - J^{(\beta)}[\rho_s, \rho_Y, 2\sqrt{\lambda}],
$$

where $\rho_s = \rho_S$ is the asymptotic density of singular values of $\boldsymbol{s} \sim P_{S,N}$. We deduce from (51) that the minimum mean-square error verifies

$$
\begin{aligned}
&\frac{1}{N^2} \mathbb{E}\|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2 \\
&= \frac{2}{N} \mathbb{E}\mathrm{Tr}\boldsymbol{\sigma}_s^2 - \frac{2}{\beta} \frac{d}{d\lambda} J_N^{(\beta)}(\boldsymbol{\sigma}^s, \boldsymbol{\sigma}^Y, \sqrt{\lambda}) + o_N(1),
\end{aligned}
$$

or, working with asymptotic densities of singular values,

$$
\begin{aligned}
&\frac{1}{N^2} \mathbb{E}\|\boldsymbol{S} - \mathbb{E}[\boldsymbol{S}|\boldsymbol{Y}]\|^2 \\
&\to 2 \int d\rho_s(x) \, x^2 - \frac{2}{\beta} \frac{d}{d\lambda} J^{(\beta)}[\rho_s, \rho_Y, \sqrt{\lambda}].
\end{aligned}
$$

V. HERMITIAN POSITIVE DEFINITE DICTIONARY LEARNING

We now move to the more challenging problem of dictionary learning, first in the positive-definite case. Its analysis will require the introduction of the main methodological advancement of this paper: the spectral replica method.

A. The model

Consider a ground-truth matrix signal $\boldsymbol{X} = [X_{ik}] \in \mathbb{K}^{N \times M}$, with

$$
M = \alpha N + o(N)
$$

with fixed $\alpha > 0$, and with prior distribution $\boldsymbol{X} \sim P_{X,N}$ which is centered $\mathbb{E} X_{ik} = 0$ and such that typically $X_{ik} = O(1)$. This

prior is *not* necessarily rotationally invariant nor factorized over the matrix entries, but we require that it induces a permutation symmetric j.p.d.f. over the singular values of X. Let $Z = Z^\dagger \in \mathbb{K}^{N \times N}$ a noise Wigner matrix with p.d.f.

$$P_{Z,N}(Z) \propto \exp \mathrm{Tr}\left(-\frac{\beta}{4}Z^2\right).$$

With this scaling the eigenvalues of Z are $O(\sqrt{N})$. Consider having access to an Hermitian data matrix $Y = [Y_{ij}] \in \mathbb{K}^{N \times N}$ with entries generated through the following observation channel:

$$Y = \sqrt{\frac{\lambda}{N}} XX^\dagger + Z. \tag{52}$$

Matrix $\sqrt{\lambda/N}\, XX^\dagger$ has $O(\sqrt{N})$ eigenvalues like the noise [and generically $O(1)$ entries], hence the scaling $\sqrt{\lambda/N}$ of the signal-to-noise ratio. The Bayesian posterior reads

$$dP_{X|Y,N}(x|Y)$$
$$= \frac{1}{\mathcal{Z}(Y)} dP_{X,N}(x) \exp \frac{\beta}{2} \mathrm{Tr}\left[\sqrt{\frac{\lambda}{N}} Yxx^\dagger - \frac{\lambda}{2N}(xx^\dagger)^2\right].$$

Note the invariance of the model under $X \to XU$ for any $M \times M$ orthogonal or unitary U such that $P_{X,N}(XU) = P_{X,N}(X)$. The mutual information $I(Y;X)$, which we aim at computing, is obtained by similar manipulations as in the previous sections:

$$\frac{1}{MN}I(Y;X) = -\frac{1}{MN}\mathbb{E}_Y \ln \int dP_{X,N}(x)$$
$$\times \exp \frac{\beta}{2}\mathrm{Tr}\left[\sqrt{\frac{\lambda}{N}} Yxx^\dagger - \frac{\lambda}{2N}(xx^\dagger)^2\right]$$
$$+ \frac{\beta\lambda}{4MN^2}\mathbb{E}\mathrm{Tr}(XX^\dagger)^2,$$

where the first term is minus the expected free entropy

$$\mathbb{E}f_N := \frac{1}{MN}\mathbb{E}_Y \ln \mathcal{Z}(Y).$$

Note that compared with our analysis on matrix denoising, we now do consider the expectation over the data in the free entropy. In the case where X is rotationally invariant, and therefore $XX^\dagger$ too, the results of the previous section on denoising can be applied. But it is important to notice right away that even in this case, *the previous conjectures do not give any information about one of the main quantities of interest in order to evaluate the inference performance, namely, the scalar overlap between the ground-truth X and a sample x from the* posterior $P_{X|Y,N}$:

$$q := \lim_{N \to +\infty} \frac{1}{N^2}\mathbb{E}\langle|\mathrm{Tr}xX^\dagger|\rangle. \tag{53}$$

The absolute value is needed because Y contains no information about the sign of X, so x and $-x$ have same posterior weight. Only the MMSE on the product $XX^\dagger$ is obtainable through this approach, through the I-MMSE identity. But this quantity is much less interesting than q as it does not carry information about the reconstruction of the internal structure of $XX^\dagger$. In particular, as noted in [56], in the present linear-rank regime of Hermitian dictionary learning with a factorized prior $P_{X,N} = g^{\otimes N(N+1)/2}$ over the matrix entries, the MMSE on $XX^\dagger$ is expected to be a universal quantity independent of the specific distribution g of the individual entries of X (as long as the first few moments exist). This is because the entries of $XX^\dagger$ are sums of many independent random contributions and thus the resulting matrix should behave at the level of the mutual information and MMSE on $XX^\dagger$ as a random matrix from the Wishart ensemble (i.e., as if X had i.i.d. standard normal entries) due to strong universality properties [112]. Therefore, it is crucial to do the following:

(1) Access the (*a priori*) nonuniversal $P_{X,N}$-dependent scalar overlap q, both in rotationally invariant and non rotationally invariant models

(2) Go beyond models with factorized distributions over the components of the hidden matrices.

Concerning the second point: as will become clear, the spectral replica method presented below does not *a priori* require the hidden matrix X (or S, T in the nonsymmetric case) to have independent entries. But the possibility to concretely evaluate expressions in the ensuing conjectures depends on the solution of a classical (but in general highly nontrivial) RMT subproblem, namely, that of evaluating the j.p.d.f. of the matrix product between two i.i.d. samples from the prior $P_{X,N}$. As a consequence, in situations where this task can be solved (despite the lack of independence of the signal matrix entries) then quantitative predictions may be reachable. Advancing on the above two points is the main role of the spectral replica method as compared with the pure RMT approaches. To fix ideas let us consider at the moment the complex case $\beta = 2$. Model (52) is equivalent to three coupled real models:

$$\mathrm{Re}Y_{ij} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}}\frac{\langle X_i, \bar{X}_j \rangle + \langle \bar{X}_i, X_j \rangle}{2}, \frac{1}{2}\right) \text{ for } i < j \in [N]^2,$$

$$\mathrm{Im}Y_{ij} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}}\frac{\langle X_i, \bar{X}_j \rangle - \langle \bar{X}_i, X_j \rangle}{2i}, \frac{1}{2}\right) \text{ for } i < j \in [N]^2,$$

$$Y_{ii} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}}\|X_i\|^2, 1\right) \text{ for } i \in [N].$$

The average free entropy then concretely reads

$$\mathbb{E}f_N = \frac{1}{NM}\mathbb{E} \ln \int dP_{X,N}(x) \exp \sum_{i \le N}\left[\sqrt{\frac{\lambda}{N}}Y_{ii}\|x_i\|^2 - \frac{\lambda}{2N}\|x_i\|^4\right] \exp 2 \sum_{i<j}^{1,N}\left[\sqrt{\frac{\lambda}{N}}\mathrm{Re}Y_{ij}\frac{\langle x_i, \bar{x}_j \rangle + \langle \bar{x}_i, x_j \rangle}{2}\right.$$
$$\left. - \frac{\lambda}{2N}\left(\frac{\langle x_i, \bar{x}_j \rangle + \langle \bar{x}_i, x_j \rangle}{2}\right)^2\right] \exp 2 \sum_{i<j}^{1,N}\left[\sqrt{\frac{\lambda}{N}}\mathrm{Im}Y_{ij}\frac{\langle x_i, \bar{x}_j \rangle - \langle \bar{x}_i, x_j \rangle}{2i} - \frac{\lambda}{2N}\left(\frac{\langle x_i, \bar{x}_j \rangle - \langle \bar{x}_i, x_j \rangle}{2i}\right)^2\right]. \tag{54}$$

B. Replica trick and spectral replica symmetry

The important difficulty is that the potential lack of rotational invriance of X and therefore of $S = XX^\dagger$ prevents the direct use of spherical integration of the rotational degrees of freedom. Moreover, we want to access the scalar overlap q, which random matrix theory alone seems not able alone to reach. But combining random matrix theory with the replica method will allow us to overcome these difficulties. The approach starts from the replica trick:

$$\lim_{N \to +\infty} \mathbb{E} f_N = \lim_{N \to +\infty} \lim_{u \to 0_+} \frac{1}{NMu} \ln \mathbb{E}\mathcal{Z}(Y)^u$$

$$= \lim_{u \to 0_+} \lim_{N \to +\infty} \frac{1}{NMu} \ln \mathbb{E}\mathcal{Z}(Y)^u. \quad (55)$$

As in standard replica theory, we assume that the u and N limits commute, and later on that the formulas we will derive for integer u can be analytically continued to real $u \to 0_+$. We therefore need to evaluate the expectation of the replicated partition function. We directly integrate the quenched Gaussian observations in (54) using the following useful formula:

$$Y \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}} f_0, \gamma\right) \Rightarrow \mathbb{E}_{Y|f_0} \prod_{a \leqslant u} \exp\gamma\left(\sqrt{\frac{\lambda}{N}} Y f_a - \frac{\lambda}{2N} f_a^2\right)$$

$$= \prod_{a < b}^{0,u} \exp\frac{\gamma\lambda}{N} f_a f_b. \quad (56)$$

In the Bayes-optimal setting the ground truth X plays a totally similar role as one additional replica, so we rename it $x^0 := X$. We set $x_i^a = (x_{ik}^a)_{k \leqslant M}$ and introduce the notation

$$\int dP_{X,N}(\{x\}_0^u) \cdots = \int_{\mathbb{R}^{MNu}} \prod_{a=0}^{u} dP_{X,N}(x^a) \cdots .$$

Define the complex-valued $M \times M$ overlap matrix

$$\boldsymbol{Q}^{ab} := \left(\frac{1}{N}\sum_{i \leqslant N} x_{ik}^a \bar{x}_{i\ell}^b\right)_{k,\ell \leqslant M} = \frac{(x^a)^{\mathsf{T}} \bar{x}^b}{N} = (\boldsymbol{Q}^{ba})^\dagger. \quad (57)$$

Then the above formula (56) applied thrice yields

$$\mathbb{E}\mathcal{Z}(Y)^u = \int dP_{X,N}(\{x\}_0^u) \prod_{a<b}^{0,u} \exp\frac{\lambda}{N}\sum_{i \leqslant N} \|x_i^a\|^2 \|x_i^b\|^2 \times \exp\frac{2\lambda}{N}\sum_{i<j}^{1,N} \frac{\langle x_i^a, \bar{x}_j^a\rangle + \langle \bar{x}_i^a, x_j^a\rangle}{2} \times \frac{\langle x_i^b, \bar{x}_j^b\rangle + \langle \bar{x}_i^b, x_j^b\rangle}{2}$$

$$\times \exp\frac{2\lambda}{N}\sum_{i<j}^{1,N} \frac{\langle x_i^a, \bar{x}_j^a\rangle - \langle \bar{x}_i^a, x_j^a\rangle}{2\mathrm{i}} \times \frac{\langle x_i^b, \bar{x}_j^b\rangle - \langle \bar{x}_i^b, x_j^b\rangle}{2\mathrm{i}}$$

$$= \int dP_{X,N}(\{x\}_0^u) \prod_{a<b}^{0,u} \exp\frac{\lambda}{N}\sum_{i,j \leqslant N}\left(\mathrm{Re}\langle x_i^a, \bar{x}_j^a\rangle \mathrm{Re}\langle x_i^b, \bar{x}_j^b\rangle + \mathrm{Im}\langle x_i^a, \bar{x}_j^a\rangle \mathrm{Im}\langle x_i^b, \bar{x}_j^b\rangle\right)$$

$$= \int dP_{X,N}(\{x\}_0^u) \prod_{a<b}^{0,u} \exp\frac{\lambda}{N}\sum_{i,j \leqslant N} \mathrm{Re}\left(\langle x_i^a, \bar{x}_j^a\rangle \langle \bar{x}_i^b, x_j^b\rangle\right)$$

$$= \int dP_{X,N}(\{x\}_0^u) \prod_{a<b}^{0,u} \exp\lambda N \sum_{k,\ell \leqslant M}\left(\frac{1}{N}\sum_{i \leqslant N} x_{ik}^a \bar{x}_{i\ell}^b\right)\left(\frac{1}{N}\sum_{j \leqslant N}\bar{x}_{jk}^a x_{j\ell}^b\right)$$

$$= \int dP_{X,N}(\{x\}_0^u) \prod_{a<b}^{0,u} \exp\lambda N \mathrm{Tr}\boldsymbol{Q}^{ab}(\boldsymbol{Q}^{ab})^\dagger. \quad (58)$$

Until now the computation is standard. Our difference starts here. Let the singular value decomposition

$$\boldsymbol{Q}^{ab} = \boldsymbol{A}^{ab}\boldsymbol{\sigma}^{ab}\boldsymbol{B}^{ab}.$$

All matrices are of size $M \times M$; note that the overlaps $(\boldsymbol{Q}^{ab})_{a<b}$ have rank equal to $n := \min(N, M)$, so $(\boldsymbol{\sigma}^{ab})_{a<b}$ have n nonzero entries on their diagonal. The dependence of the replicated system in $(x^a)_a$ is through the overlap matrices $(\boldsymbol{Q}^{ab})_{a<b}$. Changing variables for $(\boldsymbol{Q}^{ab})_{a<b}$ we have the completely generic change of density

$$dP_{X,N}(\{x\}_0^u) = dP_{(Q),M}[(\boldsymbol{Q}^{ab})_{a<b}]$$

$$= dP_{(A,B)|(\sigma),M}[(\boldsymbol{A}^{ab}, \boldsymbol{B}^{ab})_{a<b}|(\boldsymbol{\sigma}^{ab})_{a<b}]$$

$$\times dP_{(\sigma),M}[(\boldsymbol{\sigma}^{ab})_{a<b}] \quad (59)$$

for a generic conditional j.p.d.f. $P_{(A,B)|(\sigma),M}$ of the singular vectors and j.p.d.f. of singular values $P_{(\sigma),M}$. With this, and because the density over singular vectors can be integrated right away, we now have

$$\mathbb{E}\mathcal{Z}^u = \int dP_{(\sigma),M}[(\boldsymbol{\sigma}^{ab})] \prod_{a<b}^{0,u} \exp\frac{\beta}{2}\lambda N \mathrm{Tr}(\boldsymbol{\sigma}^{ab})^2. \quad (60)$$

This measure couples all matrices of singular values. It thus seems hopeless to go further without assuming some sort of simplification. We are now in position to move forward due to a type of decoupling assumption which we think is the weakest (and most natural) possible assumption allowing us to carry on computations from there. The *spectral replica symmetric ansatz* states that the replicated partition

function is dominated by configurations such that the joint law $P_{(\sigma),M}[(\sigma^{ab})_{a<b}]$ factorizes as $N, M \to +\infty$ into a product of identical laws:

Assumption (spectral replica symmetry):

$$P_{(\sigma),M}[(\sigma^{ab})_{a<b}] \to \prod_{a<b}^{0,u} \tilde{p}_M(\sigma^{ab}). \qquad (61)$$

The convergence $\to$ means that both the left- and right-hand sides weakly converge to the same asymptotic distribution as $N \to +\infty$. The j.p.d.f. $\tilde{p}_M(\sigma^{ab})$, which is shared by assumption by the different overlap matrices/pairs of replica indices $a < b$, corresponds to the j.p.d.f. of the singular values of the overlap $N^{-1}(x^a)^{\mathsf{T}} \bar{x}^b$ [or equivalently $N^{-1}(x^b)^\dagger x^a$] between two i.i.d. matrices x^a, x^b drawn from the prior $P_{X,N}$, and can thus in theory be obtained using random matrix theory; we will discuss further this point in the next section. Denoting $\sigma^{ab} = \sigma^{ab}(x^a, x^b)$ the singular values of the overlap,

$$\tilde{p}_M(\sigma) = \int dP_{X,N}(x^a) dP_{X,N}(x^b) \delta(\sigma^{ab} - \sigma). \qquad (62)$$

The law $\tilde{p}_M$ enforces $M - \min(N, M)$ singular values to be identically zero. We denote by p_n the law of the remaining n nonzero singular values only, that we continue to denote σ^{ab} with a slight abuse of notation. Note that we did *not* assume anything at the level of singular vectors which is an important feature. The decoupling is only assumed at the spectral level. So we have now that at leading exponential order and as N gets large,

$$\mathbb{E}\mathcal{Z}^u = \int \prod_{a<b}^{0,u} d\tilde{p}_M(\sigma^{ab}) \exp \frac{\beta}{2} \lambda N \mathrm{Tr}(\sigma^{ab})^2$$

$$= \int \prod_{a<b}^{0,u} dp_n(\sigma^{ab}) \exp \frac{\beta}{2} \lambda N \mathrm{Tr}(\sigma^{ab})^2$$

$$= \left[\int dp_n(\sigma) \exp \frac{\beta}{2} \lambda N \mathrm{Tr}\sigma^2 \right]^{u(u+1)/2}. \qquad (63)$$

Therefore the replica computation gives, using formula (55), that the expected free entropy limit is given by

$$\lim_{N \to +\infty} \frac{1}{2MN} \ln \int d\sigma \exp MN \left(\frac{\ln p_n(\sigma)}{MN} + \frac{\beta\lambda}{2M} \mathrm{Tr}\sigma^2 \right),$$

which can be estimated by saddle point.

C. Replica symmetric formula

Our computation shows that the order parameters are the (nonzero) singular values σ of the overlap matrix $Q := N^{-1} x^\dagger \tilde{x}$ between i.i.d. samples $x, \tilde{x}$ from the posterior distribution $P_{X|Y,N}$ (i.e., two conditionally independent replicas). Let us see how the scalar overlap q defined by (53) can be deduced from it. A general Nishimori identity reads (see [76])

$$\mathbb{E}\langle g(x, \tilde{x}) \rangle = \mathbb{E}\langle g(x, X) \rangle. \qquad (64)$$

When two or more replicas appear inside a Gibbs bracket $\langle \cdot \rangle$ it has to be understood as the expectation with respect to the product Gibbs measure. From it, one can deduce the

nonuniversal scalar overlap q. Indeed, the latter is equal to

$$q := \lim_{N \to +\infty} \frac{1}{N^2} \mathbb{E}\langle |\mathrm{Tr}xX^\dagger| \rangle = \lim_{N \to +\infty} \frac{1}{N} \mathbb{E}\langle |\mathrm{Tr}Q| \rangle$$

$$= \lim_{N \to +\infty} \frac{1}{N} |\mathrm{Tr}\mathbb{E}\langle Q \rangle| = \frac{1}{N} |\mathrm{Tr}Q| + o_N(1). \qquad (65)$$

The second equality follows from (64), while the third and last by concentration of the spectral moments of Q; this does not mean that Q concentrates elementwise, only the moments $N^{-1}\mathrm{Tr}Q^k = N^{-1}\mathrm{Tr}\mathbb{E}\langle Q \rangle^k + o_N(1)$ do. Finally, because $\mathbb{E}\langle Q \rangle = \mathbb{E}[\langle x \rangle \langle x \rangle^\dagger]$ is positive definite, its trace is also the sum of its singular values which, by the assumed self-averaging of the (moments of the) distribution of singular values, must be relatively close to $\mathrm{Tr}\sigma$ of singular values of Q (which is not symmetric). Therefore, the mean of the singular values of the overlap matrix yield the scalar overlap q.

Result 5 (Replica symmetric formula for Hermitian dictionary learning). Let p_n be the j.p.d.f. of the $n := \min(M, N)$ nonzero singular values (which are $O(1)$) of the matrix $N^{-1}x^\dagger \tilde{x}$, where $x, \tilde{x}$ are i.i.d. $N \times M$ (not necessarily rotationally invariant) random matrices drawn from the prior $P_{X,N}$. The mutual information of model (52) verifies

$$\frac{1}{MN} I\left(X; \sqrt{\frac{\lambda}{N}} XX^\dagger + Z \right)$$

$$= -\frac{1}{2} \sup_{\sigma \in \mathbb{R}^n_{\geqslant 0}} \left\{ \frac{\ln p_n(\sigma)}{MN} + \frac{\beta\lambda}{2M} \mathrm{Tr}\sigma^2 \right\} + \frac{\beta\lambda}{4} \frac{\mathbb{E}\mathrm{Tr}(XX^\dagger)^2}{MN^2}$$

$$+ \frac{1}{2} \sup_{\sigma \in \mathbb{R}^n_{\geqslant 0}} \frac{\ln p_n(\sigma)}{MN} + o_N(1). \qquad (66)$$

Denote σ^* the overlap singular values achieving the supremum in (66). The overlap (65) is then

$$q = \frac{1}{N} \mathrm{Tr}\sigma^* + o_N(1). \qquad (67)$$

D. Validity of the spectral replica method in a simple scenario

Let us show that Result 5 is correct in a simple (but nontrivial) setting. Let $N = M = n$. Consider the case where $X = \sqrt{N}O$ for an orthogonal ($\beta = 1$) or unitary ($\beta = 2$) matrix O. Then $XX^\dagger$ is N times the identity and, therefore, the mutual information is null (the data Y being independent of X). The law p_n in Result 5 thus simply enforces all singular values to be one for any λ. Thus (66) becomes

$$0 = -\frac{\beta\lambda}{4N} \mathrm{Tr}\sigma^2 + \frac{\beta\lambda}{4} \frac{\mathbb{E}\mathrm{Tr}(XX^\dagger)^2}{N^3}$$

with all $\sigma_i = 1$, which is indeed true.

E. Working with a density order parameter, and a weaker spectral replica symmetric ansatz

Let us show that the variational problem in Result 5 can equivalently be expressed as an optimization over a density, similarly as in Sec. II C. For the rest of this section we consider that the prior $P_{X,N}$ defines a rotationally invariant ensemble of random matrices in the sense that matrices of left and right singular vectors of $X \sim P_{X,N}$ are Haar distributed. Another interest of the following derivation is that it will be based on

a weaker form of the spectral replica symmetric ansatz (61). We start again from the averaged replicated partition function (60). Because the exponents are functions of the nonzero singular values only, we can simply think of $P_{(\sigma),M}[(\sigma^{ab})]$ entering (60) as the j.p.d.f. of these n nonzero singular values. The empirical distribution $n^{-1}\sum_{k\leqslant n}\delta[(\sigma_k^{ab})^2 - x]$ of the nonzero squared singular values of the overlap Q^{ab} is also the empirical distribution $\hat{\rho}_n^{ab}$ of the nonzero eigenvalues $(\lambda_i^{ab})_{i\leqslant n}$ of the nonnegative definite Hermitian matrix $Q^{ab}(Q^{ab})^\dagger$:

$$\hat{\rho}_n^{ab}(x) := \frac{1}{n}\sum_{i\leqslant n}\delta(\lambda_i^{ab} - x).$$

The j.p.d.f. $P_{(\sigma),M}((\sigma^{ab})_{a<b})$ induces a joint probability density $\hat{P}_n[(\hat{\rho}^{ab})_{a<b}]$ for the $u(u+1)/2$-dimensional random vector of empirical spectral measures $(\hat{\rho}_n^{ab})_{a<b}$. Then the integral (60) can be reexpressed as

$$\int D\hat{P}_n[(\hat{\rho}^{ab})_{a<b}]\prod_{a<b}^{0,u}\exp\left[\frac{\beta\lambda}{2}NM\int d\hat{\rho}^{ab}(x)x\right]. \quad (68)$$

We *assume* that an LDP at scale $M^2 = \Theta(N^2)$ with rate functional $\tilde{L}$ holds for $(\hat{\rho}_n^{ab})_{a<b}$; this is quite generally correct for ESDs of (single) rotationally invariant random matrices [96–98], which is the reason why we assume this invariance here. This LDP reads

$$D\hat{P}_n[(\hat{\rho}^{ab})_{a<b}] \sim \exp\big\{-M^2\tilde{L}[(\hat{\rho}^{ab})_{a<b}]\big\}\prod_{a<b}^{0,u}D[\hat{\rho}^{ab}].$$

The weaker form of the spectral replica symmetric ansatz (61) that we use here states that the rate functional $\tilde{L}$ is actually additive over the pairs of indices $a < b$ or, said differently, that the ESDs $(\hat{\rho}_n^{ab})_{a<b}$ asymptotically decouple with same laws dictated by a common rate function L:

Assumption (spectral replica symmetry, weaker form):

$$\tilde{L}[(\hat{\rho}^{ab})_{a<b}] \to \sum_{a<b}^{0,u}L[\hat{\rho}^{ab}]. \quad (69)$$

This ansatz is indeed weaker than (61) as decoupling is assumed for a much coarser statistics than the j.p.d.f. of the singular values, namely, decoupling of the ESDs over the replica pairs, which capture only the one-point marginals. The rate function L is naturally taken to be the one corresponding to the ESD of $Q^{ab}(Q^{ab})^\dagger$ where x^a and x^b are now simply two i.i.d. samples from the prior $P_{X,N}$, thus L can be obtained, at least in certain cases as we will discuss in the next section. Therefore (68) simplifies into

$$\left\{\int D[\hat{\rho}]\exp\left[\frac{\beta\lambda}{2}NM\int d\hat{\rho}(x)x - M^2L[\hat{\rho}]\right]\right\}^{u(u+1)/2}.$$

Plugging this expression in (55) and evaluating the integral by saddle point over the density gives an expression in the case where the prior $P_{X,N}$ generates rotationally invariant matrices.

Result 6 (Replica symmetric formula for Hermitian dictionary learning with rotationally invariant prior). Consider the rotationally invariant random matrix ensemble defined by Hermitian matrices $T := N^{-2}x^\dagger\tilde{x}\tilde{x}^\dagger x$ where $x, \tilde{x}$ are i.i.d. $N \times M$ samples from the rotationally invariant prior $P_{X,N}$.

Assume that the empirical spectral distribution of the nonzero eigenvalues of T verifies a LDP at scale M^2 with rate functional L. The asymptotic mutual information of model (52) then verifies

$$\frac{1}{MN}I\left(X;\sqrt{\frac{\lambda}{N}}XX^\dagger + Z\right)$$

$$\to -\frac{1}{2}\sup_{\rho\in\mathcal{P}_{\geqslant 0}}\left\{\frac{\beta\lambda}{2}\int d\rho(x)x - \alpha L[\rho]\right\}$$

$$+ \frac{1}{2}\sup_{\rho\in\mathcal{P}_{\geqslant 0}}\{-\alpha L[\rho]\} + \lim_{N\to+\infty}\frac{\beta\lambda}{4}\frac{\mathbb{E}\operatorname{Tr}(XX^\dagger)^2}{MN^2}. \quad (70)$$

Denote ρ^* the supremum in (70), which corresponds to the asymptotic density of nonzero eigenvalues of the (expected) overlap matrix times its conjugate transpose. Using a standard change of density we can access the asymptotic density $\gamma^* \in \mathcal{P}_{\geqslant 0}$ of singular values of the overlap using $\gamma^*(x) = 2x\rho^*(x^2)$. Let $a := \min(1, \alpha)$. The overlap (65) is then

$$q = \frac{1}{a}\int d\gamma^*(x)x = \frac{1}{a}\int d\rho^*(x)\sqrt{x}. \quad (71)$$

In case the rate functional L' for the LDP of the empirical density of nonzero singular values of the matrix $N^{-1}x^\dagger\tilde{x}$ is easier to obtain than L, then formula (70) is simply changed by replacing L by L' as well as the expectation $\int d\rho(x)x$ by the second moment $\int d\rho(x)x^2$. The optimizer ρ^* of the resulting formula then gives access to the scalar overlap due to $q = \int d\rho^*(x)x/a$. Assume there exists an effective potential V_T underlying this ensemble, in the sense that $T' \sim \exp[-\frac{\beta M}{4}\operatorname{Tr}V_T(T')]$ has same asymptotic spectral density as T. This effective potential V_T can be deduced from the asymptotic spectral density using (21). Then by approximation of the rate functional L by the usual form (18) for rotationally invariant ensembles, the asymptotic mutual information can be approximated by

$$\frac{1}{MN}I\left(X;\sqrt{\frac{\lambda}{N}}XX^\dagger + Z\right)$$

$$\approx -\frac{1}{2}\sup_{\rho\in\mathcal{P}_{\geqslant 0}}\left[\frac{\beta\lambda}{2}\int d\rho(x)x - \frac{\alpha\beta}{4}\int d\rho(x)V_T(x)\right.$$

$$\left. + \frac{\alpha\beta}{2}\int d\rho(x)\,d\rho(y)\ln|x - y|\right]$$

$$+ \frac{1}{2}\sup_{\rho\in\mathcal{P}_{\geqslant 0}}\left[\frac{\alpha\beta}{2}\int d\rho(x)\,d\rho(y)\ln|x - y|\right.$$

$$\left. - \frac{\alpha\beta}{4}\int d\rho(x)V_T(x)\right] + \lim_{N\to+\infty}\frac{\beta\lambda}{4}\frac{\mathbb{E}\operatorname{Tr}(XX^\dagger)^2}{MN^2}. \quad (72)$$

F. A more explicit special case: The Ginibre signal

Obtaining the j.p.d.f. p_n of singular values entering Result 5 from the knowledge of $P_{X,N}$ or the rate functional L in Result 6 may be highly nontrivial, but this has the merit of being well-defined "standard" random matrix theory problems. In certain cases these are known. For example, for products of i.i.d. Gaussian (Ginibre) matrices (or with an additional source [113]). Starting with the j.p.d.f., in this case the law p_n is related to the partition function of a two-matrix model (see

next section) that can be integrated and yields a determinant point process defined in terms of the Meijer G function; see [94,114–118]. Products of finitely but arbitrarily many matrices are considered in these references, but for us only the case of a product between two matrices is needed. There also exist results for products of truncated unitary matrices [119,120] and for more general product ensembles (but with less explicit formulas) [95]. We refer to [121] for a review on the subject. See also [122,123] for information about fluctuations and universality properties in such matrix product ensembles, or [114,124–127] for results concerning their asymptotic density of eigenvalues and singular values. Using this body of work we can go further in explicating Results 5 and 6 in these known cases. In this section we restrict ourselves to the special case where the signal X is a Ginibre matrix, so that $XX^\dagger$ is Wishart. We want to find the j.p.d.f. of the singular values of the matrix $N^{-1}x_0^\dagger \tilde{x}_0$ where $x_0, \tilde{x}_0$ are i.i.d. $N \times M$ Ginibre matrices with $O(1)$ entries, which is equivalent to find it for the matrix $y_0^\dagger \tilde{y}_0$ where $y_0, \tilde{y}_0$ are i.i.d. $N \times M$ standard Ginibre matrices with law

$$P_{Y_0,N}(y_0) \propto \exp \mathrm{Tr}\left(-\frac{\beta N}{2}y_0 y_0^\dagger\right).$$

Recall $n := \min(N, M)$. The steps leading to (2.11) of [118] for the square case $M = N$, or those leading to (16) of [94] for the general rectangular case (where M and N do not necessarily match), imply that the j.p.d.f. p_n of the n nonzero singular values σ of the matrix $y_0^\dagger \tilde{y}_0$ is

$$p_n(\sigma) \propto |\Delta_n(\sigma^2)|^\beta \left(\prod_{k \leqslant n} \sigma_k\right)^{\beta(M-n+1)-1} R_n(\sigma)$$

$$\propto \exp MN\left[\frac{\beta}{MN}\sum_{i<j}^{1,n} \ln|\sigma_i^2 - \sigma_j^2| + \beta\frac{M-n}{MN}\mathrm{Tr}\ln\sigma\right.$$

$$\left. + \frac{1}{MN}\ln R_n(\sigma) + o_n(1)\right], \tag{73}$$

where the function

$$R_n(\sigma) := \int_{\mathbb{R}_{\geqslant 0}^n} d\boldsymbol{r}\, |\Delta_n(r^2)|^\beta \exp\mathrm{Tr}\left(-\frac{\beta N}{2}r^2\right)$$

$$\times \int d\mu_n^{(\beta)}(U) \exp\mathrm{Tr}\left(-\frac{\beta N}{2}U^\dagger \sigma^2 U r^{-2}\right)$$

$$\times \left(\prod_{k \leqslant n} r_k\right)^{\beta(N-M-n+1)-1}. \tag{74}$$

We recognize a two-matrix model. The spherical integral appears in the function $R_n(\sigma)$, that we reexpress in a form appropriate for a saddle-point evaluation:

$$\int d\boldsymbol{r} \exp MN\left[\frac{\beta}{MN}\sum_{i<j}^{1,n}\ln|r_i^2 - r_j^2| - \frac{\beta}{2M}\mathrm{Tr}\,r^2\right.$$

$$\left. + \beta\frac{N-M-n}{MN}\mathrm{Tr}\ln r + I_n^{(\beta)}\left(\sigma^2, r^{-2}, -\frac{N}{n}\right) + o_n(1)\right].$$

Therefore we reach

$$\frac{\ln R_n(\sigma)}{MN} = \sup_{\boldsymbol{r}\in\mathbb{R}_{\geqslant 0}^n}\left[\frac{\beta}{MN}\sum_{i<j}^{1,n}\ln|r_i^2 - r_j^2| - \frac{\beta}{2M}\mathrm{Tr}\,r^2\right.$$

$$\left. + \beta\frac{N-M-n}{MN}\mathrm{Tr}\ln r + I_n^{(\beta)}\left(\sigma^2, r^{-2}, -\frac{N}{n}\right)\right]$$

$$+ o_n(1).$$

Thus we end up with

$$\frac{\ln p_n(\sigma)}{MN}$$

$$= \beta\sup_{\boldsymbol{r}\in\mathbb{R}_{\geqslant 0}^n}\left[\sum_{i\neq j}^{1,n}\frac{\ln|\sigma_i^2 - \sigma_j^2|}{2MN} + \frac{M-n}{2MN}\mathrm{Tr}\ln\sigma^2\right.$$

$$+ \sum_{i\neq j}^{1,n}\frac{\ln|r_i^2 - r_j^2|}{2MN} - \frac{\mathrm{Tr}\,r^2}{2M} + \frac{N-M-n}{2MN}\mathrm{Tr}\ln r^2$$

$$\left. + \frac{n^2}{\beta MN}I_n^{(\beta)}\left(\sigma^2, r^{-2}, -\frac{N}{n}\right)\right] + o_n(1). \tag{75}$$

With this expression in hand we can write down a more explicit version of Result 5 when the signal is Ginibre. By the same type of manipulations as those found in Sec. II C we can also identify the rate functional L starting from (75) and therefore express the result in terms of a density order parameter.

Result 7 (Replica symmetric formula for Hermitian dictionary learning with a Ginibre signal). Let $n := \min(N, M)$. When $P_{X,N}(X) \propto \exp\mathrm{Tr}[-(\beta/2)XX^\dagger]$ the mutual information of model (52) verifies

$$\frac{1}{MN}I\left(X; \sqrt{\frac{\lambda}{N}}XX^\dagger + Z\right)$$

$$= -\frac{\beta}{2}\sup_{(\sigma,\boldsymbol{r})\in\mathbb{R}_{\geqslant 0}^n \times \mathbb{R}_{\geqslant 0}^n}\left[\sum_{i\neq j}^{1,n}\frac{\ln|\sigma_i^2 - \sigma_j^2|}{2MN} + \sum_{i\neq j}^{1,n}\frac{\ln|r_i^2 - r_j^2|}{2MN}\right.$$

$$+ \frac{M-n}{2MN}\mathrm{Tr}\ln\sigma^2 + \frac{N-M-n}{2MN}\mathrm{Tr}\ln r^2$$

$$+ \frac{\lambda}{2M}\mathrm{Tr}\sigma^2 - \frac{1}{2M}\mathrm{Tr}\,r^2 + \frac{n^2}{\beta MN}I_n^{(\beta)}\left(\sigma^2, r^{-2}, -\frac{N}{n}\right)\right]$$

$$+ \frac{\beta\lambda}{4}\frac{\mathbb{E}\mathrm{Tr}(XX^\dagger)^2}{MN^2} + \tau_N. \tag{76}$$

Constant τ_N fixes $I(X; Z) = 0$:

$$\tau_N := \frac{\beta}{2}\sup_{(\sigma,\boldsymbol{r})\in\mathbb{R}_{\geqslant 0}^n \times \mathbb{R}_{\geqslant 0}^n} g_n^{\mathrm{RS}}(\sigma, \boldsymbol{r}, \lambda = 0) + o_N(1)$$

for the replica symmetric potential function $g_n^{\mathrm{RS}} : \mathbb{R}_{\geqslant 0}^n \times \mathbb{R}_{\geqslant 0}^n \times \mathbb{R}_{\geqslant 0} \mapsto \mathbb{R}$ defined by the curly bracket $\{\cdots\}$ in the variational problem (76). Denote σ^* the overlap singular values achieving the supremum in (76). The overlap (65) is then given by $q = N^{-1}\mathrm{Tr}\sigma^* + o_N(1)$. Introducing asymptotic densities ρ and $\tilde{\rho}$ associated to the squared singular values σ^2 and r^2, respectively, the conjecture can be reexpressed in the

limit $N \rightarrow +\infty$ and $M/N \rightarrow \alpha$ as

$$
\frac{1}{MN}I\left(\boldsymbol{X}; \sqrt{\frac{\lambda}{N}}\boldsymbol{X}\boldsymbol{X}^{\dagger} + \boldsymbol{Z}\right)
$$

$$
\rightarrow -\frac{\beta}{2}\sup_{(\rho,\tilde{\rho})\in\mathcal{P}_{\geqslant 0}\times\mathcal{P}_{\geqslant 0}}\left\{\min\left(\frac{\alpha}{2}, \frac{1}{2\alpha}\right)\left[\int d\rho(x)\,d\rho(y)\ln|x-y| + \int d\tilde{\rho}(x)\,d\tilde{\rho}(y)\ln|x-y|\right]\right.
$$

$$
+\max\left(0, \frac{1}{2} - \frac{1}{2\alpha}\right)\int d\rho(x)\ln x + \max\left(\frac{1}{2} - \alpha, -\frac{1}{2}\right)\int d\tilde{\rho}(x)\ln x + \min\left(\frac{1}{2}, \frac{1}{2\alpha}\right)\left[\lambda\int d\rho(x)x - \int d\tilde{\rho}(x)x\right]
$$

$$
\left. +\frac{1}{\beta}\min\left(\alpha, \frac{1}{\alpha}\right)I^{(\beta)}\left[\rho, \hat{\rho}[\tilde{\rho}], \frac{-1}{\min(1,\alpha)}\right]\right\} + \frac{\beta\lambda}{4\alpha}\lim_{N\rightarrow+\infty}\frac{\mathbb{E}\operatorname{Tr}(\boldsymbol{X}\boldsymbol{X}^{\dagger})^2}{N^3} + \tau. \tag{77}
$$

Using a standard change of density the density $\hat{\rho}[\tilde{\rho}]$ belonging to $\mathcal{P}_{\geqslant 0}$ can be expressed in terms of $\tilde{\rho}$ over which the optimization takes place:

$$
\hat{\rho}[\tilde{\rho}](x) = \frac{\tilde{\rho}(1/x)}{x^2}.
$$

Constant τ fixes the constraint $I(\boldsymbol{X}; \boldsymbol{Z}) = 0$:

$$
\tau := \frac{\beta}{2}\sup_{(\rho,\tilde{\rho})\in\mathcal{P}_{\geqslant 0}\times\mathcal{P}_{\geqslant 0}}g^{\mathrm{RS}}[\rho, \tilde{\rho}, \lambda = 0]
$$

for the replica symmetric potential functional $g^{\mathrm{RS}} : \mathcal{P}_{\geqslant 0} \times \mathcal{P}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \mapsto \mathbb{R}$ defined by the curly brackets $\{\cdots\}$ in (77). Denoting by ρ^* the density of squared singular values of the overlap that achieves the supremum over ρ in (77), the scalar overlap reads $q = \int d\rho^*(x)\sqrt{x}/\min(1,\alpha)$.

Due to the presence of the spherical integral, solving the above variational problem(s) is an important challenge. Because the Ginibre prior induces a rotationally invariant (Wishart) matrix $\boldsymbol{X}\boldsymbol{X}^{\dagger}$ this formula must match the simpler Result 3. But, as already mentioned, the great advantage of Result 7 (and more generically of Results 5 and 6) is that their numerical solution would also give access to the scalar overlap q, which cannot be extracted from Result 3. Solving the functional variational problems of Result 7 and comparing it to Result 3, whose solution can be numerically obtained thanks to Matytsin's formalism [88], is therefore one (rather challenging) route to check the validity of our spectral replica method. We do not know at the moment how to do that, and this is left for future investigations. But there is another alternative route to check the validity of our replica theory without actually solving the above variational problems. It goes as follows. Recall the integral appearing in (63). In the case of a Ginibre signal it reads

$$
\int d\boldsymbol{\sigma}\,d\boldsymbol{r}\,|\Delta_n(\boldsymbol{\sigma}^2)\Delta_n(\boldsymbol{r}^2)|^{\beta}
$$

$$
\times\left(\prod_{k\leqslant n}\sigma_k\right)^{\beta(M-n+1)-1}\left(\prod_{k\leqslant n}r_k\right)^{\beta(N-M-n+1)-1}
$$

$$
\times\int d\mu_n^{(\beta)}(\boldsymbol{U})\exp\frac{\beta N}{2}\operatorname{Tr}(\lambda\boldsymbol{\sigma}^2 - \boldsymbol{r}^2 - \boldsymbol{U}^{\dagger}\boldsymbol{\sigma}^2\boldsymbol{U}\boldsymbol{r}^{-2}).
$$

This is a two-matrix model. If one was able to actually compute this integral without using a saddle point (which generates the variational formulation), the result could be compared to Result 3. Fortunately, it can be computed exactly

using the method of biorthogonal polynomials as used in [66,94,118]; see also [128]. This promising strategy to test the validity of our spectral replica method is nontrivial and left for future work.

VI. DICTIONARY LEARNING

A. The model

Let the ground-truth dictionary $\boldsymbol{S} = [S_{ik}] \in \mathbb{K}^{N\times K}$ be drawn from a centered distribution $\boldsymbol{S} \sim P_{S,K}$, and the coefficients $\boldsymbol{T} = [T_{jk}] \in \mathbb{K}^{M\times K}$ from $\boldsymbol{T} \sim P_{T,K}$ centered also, where the entries of both $\boldsymbol{S}$ and $\boldsymbol{T}$ are typically $O(1)$. These two priors should induce symmetric j.p.d.f. of singular values for $\boldsymbol{S}$ and $\boldsymbol{T}$. We set

$$
N = \alpha K + o(K) \quad \text{and} \quad M = \gamma K + o(K)
$$

with fixed $\alpha, \gamma > 0$ as $K \rightarrow +\infty$. Consider having access to a data matrix $\boldsymbol{Y} = [Y_{ij}] \in \mathbb{K}^{N\times M}$ with entries generated according to

$$
\boldsymbol{Y} = \sqrt{\frac{\lambda}{N}}\boldsymbol{S}\boldsymbol{T}^{\dagger} + \boldsymbol{Z}, \tag{78}
$$

with a Ginibre noise matrix $\boldsymbol{Z} \in \mathbb{K}^{N\times M}$ with law

$$
P_{Z,N}(\boldsymbol{Z}) \propto \exp\operatorname{Tr}\left(-\frac{\beta}{2}\boldsymbol{Z}\boldsymbol{Z}^{\dagger}\right).
$$

The scaling in N of the matrix entries and eigenvalues are the same as in the positive definite case (52). We assume this time that both priors are biorthogonal or unitary rotationally invariant, i.e.,

$$
dP_{S,K}(\boldsymbol{S}) = dP_{S,K}(\boldsymbol{O}\boldsymbol{S}\tilde{\boldsymbol{O}}),
$$

$$
dP_{T,K}(\boldsymbol{T}) = dP_{T,K}(\boldsymbol{O}\boldsymbol{T}\tilde{\boldsymbol{O}}) \tag{79}
$$

for any orthogonal ($\beta = 1$) or unitary ($\beta = 2$) matrices $\boldsymbol{O}$ and $\tilde{\boldsymbol{O}}$. Rotational invariance of the prior was not needed in the Hermitian case of dictionary learning, but it seems required in the less symmetric present setting. We jointly denote $\boldsymbol{X} := (\boldsymbol{S}, \boldsymbol{T})$ and $\boldsymbol{x} := (\boldsymbol{s}, \boldsymbol{t}) \in \mathbb{K}^{N\times K} \times \mathbb{K}^{M\times K}$. Let $dP_{X,K}(\boldsymbol{x}) := dP_{S,K}(\boldsymbol{s})\,dP_{T,K}(\boldsymbol{t})$. The joint posterior is

$$
dP_{X|Y,K}(\boldsymbol{x}|\boldsymbol{Y}) = \frac{1}{\mathcal{Z}(\boldsymbol{Y})}dP_{X,K}(\boldsymbol{x})\exp\frac{\beta}{2}\operatorname{Tr}
$$

$$
\times\left(\sqrt{\frac{\lambda}{N}}\boldsymbol{Y}^{\dagger}\boldsymbol{s}\boldsymbol{t}^{\dagger} + \sqrt{\frac{\lambda}{N}}\boldsymbol{Y}\boldsymbol{t}\boldsymbol{s}^{\dagger} - \frac{\lambda}{N}\boldsymbol{s}^{\dagger}\boldsymbol{s}\boldsymbol{t}^{\dagger}\boldsymbol{t}\right). \tag{80}
$$

Note the invariance of the model under $(\boldsymbol{S}, \boldsymbol{T}) \to (\boldsymbol{S}\boldsymbol{U}, \boldsymbol{T}\boldsymbol{U})$ for any orthogonal or unitary $\boldsymbol{U}$. The object of interest is the average free entropy

$$\mathbb{E}f_N := \frac{1}{NM}\mathbb{E}\ln\mathcal{Z}(\boldsymbol{Y}).$$

It is linked to the mutual information by

$$\frac{1}{MN}I[\boldsymbol{Y}; (\boldsymbol{S}, \boldsymbol{T})] = -\mathbb{E}f_N + \frac{\beta\lambda}{2MN^2}\mathbb{E}\mathrm{Tr}\boldsymbol{S}^\dagger\boldsymbol{S}\boldsymbol{T}^\dagger\boldsymbol{T}.$$

B. Replica trick and spectral replica symmetry

As before, working in the Bayes-optimal setting allows us to simply rename the ground truth $\boldsymbol{x}^0 = \boldsymbol{X}$ which will play the same role as all other replicas $\boldsymbol{x}^a = (\boldsymbol{s}^a, \boldsymbol{t}^a) \in \mathbb{K}^{N \times K} \times \mathbb{K}^{M \times K}$ of $\boldsymbol{x}$. We set

$$\int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \cdots$$
$$= \int_{\mathbb{K}^{NK(u+1)}} \prod_{a=0}^u dP_{S,K}(\boldsymbol{s}^a) \int_{\mathbb{K}^{MK(u+1)}} \prod_{a=0}^u dP_{T,K}(\boldsymbol{t}^a) \cdots.$$

The replica trick (55) requires computing the moments $\mathbb{E}\mathcal{Z}^u$ of the partition function. As in Sec. V we consider first the more cumbersome complex case $\beta = 2$. Model (78) is then

equivalent to

$$\mathrm{Re}Y_{ij} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}} \frac{\langle \boldsymbol{S}_i, \bar{\boldsymbol{T}}_j\rangle + \langle \bar{\boldsymbol{S}}_i, \boldsymbol{T}_j\rangle}{2}, \frac{1}{2}\right)$$
$$(i, j) \in [N] \times [M].$$
$$\mathrm{Im}Y_{ij} \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{N}} \frac{\langle \boldsymbol{S}_i, \bar{\boldsymbol{T}}_j\rangle - \langle \bar{\boldsymbol{S}}_i, \boldsymbol{T}_j\rangle}{2\mathrm{i}}, \frac{1}{2}\right)$$

We introduce the $K \times K$ (*a priori* non-Hermitian) overlap matrices and their SVD decompositions:

$$\boldsymbol{Q}_s^{ab} := \left(\frac{1}{N}\sum_{i \leqslant N} s_{ik}^a \bar{s}_{i\ell}^b\right)_{k,\ell \leqslant K} = \frac{(\boldsymbol{s}^a)^\mathsf{T}\bar{\boldsymbol{s}}^b}{N} = \boldsymbol{A}_s^{ab}\boldsymbol{\sigma}_s^{ab}\boldsymbol{B}_s^{ab},$$

$$\boldsymbol{Q}_t^{ab} := \left(\frac{1}{N}\sum_{j \leqslant M} \bar{t}_{jk}^a t_{j\ell}^b\right)_{k,\ell \leqslant K} = \frac{(\boldsymbol{t}^a)^\dagger\boldsymbol{t}^b}{N} = \boldsymbol{A}_t^{ab}\boldsymbol{\sigma}_t^{ab}\boldsymbol{B}_t^{ab}.$$

The overlaps $(\boldsymbol{Q}_s^{ab})_{a<b}$ are of rank $n_s := \min(N, K)$, and thus $(\boldsymbol{\sigma}_s^{ab})_{a<b}$ have n_s nonzero entries on their diagonal, while $(\boldsymbol{Q}_t^{ab})_{a<b}$ have rank $n_t := \min(M, K)$ implying that $(\boldsymbol{\sigma}_t^{ab})_{a<b}$ have n_t nonzero entries on their diagonal. Now, we integrate $\boldsymbol{Y}$ which is, conditionally on $(\boldsymbol{S}, \boldsymbol{T})$, a complex Gaussian multivariate random variable, by using formula (56) and obtain that

$$\mathbb{E}\mathcal{Z}^u = \int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}^0} \prod_{a \leqslant u} \exp 2 \sum_{i,j}^{N,M} \left[\sqrt{\frac{\lambda}{N}}\mathrm{Re}Y_{ij} \frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle + \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2} - \frac{\lambda}{2N}\left(\frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle + \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2}\right)^2\right]$$

$$\times \exp 2 \sum_{i,j}^{N,M} \left[\sqrt{\frac{\lambda}{N}}\mathrm{Im}Y_{ij} \frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle - \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2\mathrm{i}} - \frac{\lambda}{2N}\left(\frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle - \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2\mathrm{i}}\right)^2\right]$$

$$= \int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \prod_{a<b}^{0,u} \exp \frac{2\lambda}{N} \sum_{i,j}^{N,M} \left(\frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle + \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2} \times \frac{\langle \boldsymbol{s}_i^b, \bar{\boldsymbol{t}}_j^b\rangle + \langle \bar{\boldsymbol{s}}_i^b, \boldsymbol{t}_j^b\rangle}{2} + \frac{\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle - \langle \bar{\boldsymbol{s}}_i^a, \boldsymbol{t}_j^a\rangle}{2\mathrm{i}} \times \frac{\langle \boldsymbol{s}_i^b, \bar{\boldsymbol{t}}_j^b\rangle - \langle \bar{\boldsymbol{s}}_i^b, \boldsymbol{t}_j^b\rangle}{2\mathrm{i}}\right)$$

$$= \int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \prod_{a<b}^{0,u} \exp \frac{2\lambda}{N} \sum_{i,j}^{N,M} \left(\mathrm{Re}\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle \mathrm{Re}\langle \boldsymbol{s}_i^b, \bar{\boldsymbol{t}}_j^b\rangle + \mathrm{Im}\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle \mathrm{Im}\langle \boldsymbol{s}_i^b, \bar{\boldsymbol{t}}_j^b\rangle\right)$$

$$= \int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \prod_{a<b}^{0,u} \exp \frac{2\lambda}{N} \sum_{i,j}^{N,M} \mathrm{Re}\left(\langle \boldsymbol{s}_i^a, \bar{\boldsymbol{t}}_j^a\rangle \langle \bar{\boldsymbol{s}}_i^b, \boldsymbol{t}_j^b\rangle\right)$$

$$= \int dP_{X,K}(\{\boldsymbol{x}\}_0^u) \prod_{a<b}^{0,u} \exp \beta\lambda N \mathrm{Re}\mathrm{Tr}\boldsymbol{Q}_s^{ab}(\boldsymbol{Q}_t^{ab})^\dagger.$$

The product form of the prior measure

$$dP_{X,K}(\{\boldsymbol{x}\}_0^u) = \prod_{a=0}^u dP_{S,K}(\boldsymbol{s}^a) dP_{T,K}(\boldsymbol{t}^a)$$

induces a j.p.d.f. factorized over the two types of overlaps:

$$dP_{(Q_s,Q_t),K}\left[(\boldsymbol{Q}_s^{ab}, \boldsymbol{Q}_t^{ab})_{a<b}\right]$$
$$= dP_{(Q_s),K}\left[(\boldsymbol{Q}_s^{ab})_{a<b}\right] dP_{(Q_t),K}\left[(\boldsymbol{Q}_t^{ab})_{a<b}\right]. \quad (81)$$

At this stage we need one additional assumption when compared with the Hermitian case of Sec. V. For each pair $a < b$

of replica indices let i.i.d. Haar matrices $\boldsymbol{U}^{ab}, \boldsymbol{V}^{ab} \sim \mu_K^{(\beta)}$ independent of everything. We *assume* the following equality in distribution in the large size limit [129], which is suggested by the combination of the independence (81) between the two types of overlaps together with the rotational invariance (79) of both priors (from which the overlap matrices $\boldsymbol{Q}_s^{ab}$ and $\boldsymbol{Q}_t^{ab}$ must inherit).

Assumption (equality in law): For any pair $a < b$:

$$\mathrm{Tr}\boldsymbol{Q}_s^{ab}(\boldsymbol{Q}_t^{ab})^\dagger \sim \mathrm{Tr}\left[\boldsymbol{U}^{ab}\boldsymbol{Q}_s^{ab}\boldsymbol{V}^{ab}(\boldsymbol{Q}_t^{ab})^\dagger\right]$$

as $M, N, K \to +\infty$ proportionally. As a consequence,

$$
\mathbb{E}\mathcal{Z}^u \propto \int dP_{(Q_s, Q_t), K}\big[\big(\boldsymbol{Q}_s^{ab}, \boldsymbol{Q}_t^{ab}\big)_{a<b}\big]
$$

$$
\times \prod_{a<b}^{0,u} \exp \beta\lambda N \mathrm{ReTr} \boldsymbol{Q}_s^{ab}\big(\boldsymbol{Q}_t^{ab}\big)^\dagger
$$

$$
= \int dP_{(Q_s), K}\big[\big(\boldsymbol{Q}_s^{ab}\big)_{a<b}\big] dP_{(Q_t), K}\big[\big(\boldsymbol{Q}_t^{ab}\big)_{a<b}\big]
$$

$$
\times \int \prod_{a<b}^{0,u} d\mu_K^{(\beta)}(\boldsymbol{U}^{ab}) \, d\mu_K^{(\beta)}(\boldsymbol{V}^{ab})
$$

$$
\times \exp \beta\lambda N \mathrm{ReTr}\big[\boldsymbol{U}^{ab}\boldsymbol{Q}_s^{ab}\boldsymbol{V}^{ab}\big(\boldsymbol{Q}_t^{ab}\big)^\dagger\big]. \qquad (82)
$$

We now change variables for the SVD decompositions of the overlaps:

$$
dP_{(Q_s), K}\big[\big(\boldsymbol{Q}_s^{ab}\big)_{a<b}\big] dP_{(Q_t), K}\big[\big(\boldsymbol{Q}_t^{ab}\big)_{a<b}\big]
$$

$$
= dP_{(A_s, B_s)|(\sigma_s), K}\big[\big(\boldsymbol{A}_s^{ab}, \boldsymbol{B}_s^{ab}\big)_{a<b}\big|\big(\boldsymbol{\sigma}_s^{ab}\big)_{a<b}\big]
$$

$$
\times dP_{(A_t, B_t)|(\sigma_t), K}\big[\big(\boldsymbol{A}_t^{ab}, \boldsymbol{B}_t^{ab}\big)_{a<b}\big|\big(\boldsymbol{\sigma}_t^{ab}\big)_{a<b}\big]
$$

$$
\times dP_{(\sigma_s), K}\big[\big(\boldsymbol{\sigma}_s^{ab}\big)_{a<b}\big] dP_{(\sigma_t), K}\big[\big(\boldsymbol{\sigma}_t^{ab}\big)_{a<b}\big].
$$

The spectral replica symmetric ansatz reads in this case:

Assumption (spectral replica symmetry):

$$
P_{(\sigma_s), K}\big[\big(\boldsymbol{\sigma}_s^{ab}\big)_{a<b}\big] \to \prod_{a<b}^{0,u} \tilde{p}_{S,K}\big(\boldsymbol{\sigma}_s^{ab}\big)
$$

$$
P_{(\sigma_t), K}\big[\big(\boldsymbol{\sigma}_t^{ab}\big)_{a<b}\big] \to \prod_{a<b}^{0,u} \tilde{p}_{T,K}\big(\boldsymbol{\sigma}_t^{ab}\big), \qquad (83)
$$

where $\tilde{p}_{S,K}$ corresponds to the j.p.d.f. of singular values of a product $N^{-1}\boldsymbol{s}^\dagger\tilde{\boldsymbol{s}}$ between i.i.d. samples from the prior $P_{S,K}$, and similarly for $\tilde{p}_{T,K}$ with samples from $P_{T,K}$. Again, $\to$ means that both sides weakly converge to the same distribution as $N, M, K \to +\infty$. Thus

$$
\mathbb{E}\mathcal{Z}^u \propto \int \prod_{a<b}^{0,u} d\tilde{p}_{S,K}\big(\boldsymbol{\sigma}_s^{ab}\big) \, d\tilde{p}_{T,K}\big(\boldsymbol{\sigma}_t^{ab}\big) \int dP_{(A_s, B_s)|(\sigma_s), K}\big[\big(\boldsymbol{A}_s^{ab}, \boldsymbol{B}_s^{ab}\big)_{a<b}\big|\big(\boldsymbol{\sigma}_s^{ab}\big)_{a<b}\big] dP_{(A_t, B_t)|(\sigma_t), K}\big[\big(\boldsymbol{A}_t^{ab}, \boldsymbol{B}_t^{ab}\big)_{a<b}\big|\big(\boldsymbol{\sigma}_t^{ab}\big)_{a<b}\big]
$$

$$
\times \int \prod_{a<b}^{0,u} d\mu_K^{(\beta)}(\boldsymbol{U}^{ab}) \, d\mu_K^{(\beta)}(\boldsymbol{V}^{ab}) \exp \beta\lambda N \mathrm{ReTr}\big[\boldsymbol{U}^{ab}\boldsymbol{A}_s^{ab}\boldsymbol{\sigma}_s^{ab}\boldsymbol{B}_s^{ab}\boldsymbol{V}^{ab}\big(\boldsymbol{B}_t^{ab}\big)^\dagger \boldsymbol{\sigma}_t^{ab}\big(\boldsymbol{A}_t^{ab}\big)^\dagger\big].
$$

The mechanism here is to absorb the left and right singular vectors into the Haar distributed matrices: we redefine $(\boldsymbol{A}_t^{ab})^\dagger\boldsymbol{U}^{ab}\boldsymbol{A}_s^{ab} \to \boldsymbol{U}^{ab}$ and $\boldsymbol{B}_s^{ab}\boldsymbol{V}^{ab}(\boldsymbol{B}_t^{ab})^\dagger \to \boldsymbol{V}^{ab}$ which remain independent and Haar distributed; these changes have unit Jacobian determinant. Thus, the distributions of singular vectors, which *a priori* couple the different pairs of replicas, are integrated and decoupling of the integrals over the pairs of indices $a < b$ takes place. This yields

$$
\mathbb{E}\mathcal{Z}(\boldsymbol{Y})^u \propto \left[\int d\tilde{p}_{S,K}(\boldsymbol{\sigma}_s) \, d\tilde{p}_{T,K}(\boldsymbol{\sigma}_t) d\mu_K^{(\beta)}(\boldsymbol{U}) \, d\mu_K^{(\beta)}(\boldsymbol{V}) \exp \beta\lambda N \mathrm{ReTr}\big(\boldsymbol{U}\boldsymbol{\sigma}_s\boldsymbol{V}\boldsymbol{\sigma}_t\big) \right]^{u(u+1)/2}
$$

$$
= \left\{ \int d\boldsymbol{\sigma}_s \, d\boldsymbol{\sigma}_t \exp MN\left[\frac{\ln \tilde{p}_{S,K}(\boldsymbol{\sigma}_s)}{MN} + \frac{\ln \tilde{p}_{T,K}(\boldsymbol{\sigma}_t)}{MN} + \frac{K^2}{MN} J_K^{(\beta)}(\boldsymbol{\sigma}_s, \boldsymbol{\sigma}_t, 2\lambda) \right] \right\}^{u(u+1)/2}.
$$

Saddle-point estimation and taking $u \to 0_+$ yields the following conjecture for the mutual information and, moreover, the nonuniversal scalar overlaps (see the justifications below (65)):

$$
q_s := \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}\langle|\mathrm{Tr}\boldsymbol{s}\boldsymbol{S}^\dagger|\rangle = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}\langle|\mathrm{Tr}\boldsymbol{Q}_s|\rangle
$$

$$
= \lim_{N\to+\infty} \frac{1}{N} |\mathrm{Tr}\mathbb{E}\langle\boldsymbol{Q}_s\rangle| = \frac{1}{N}|\mathrm{Tr}\boldsymbol{Q}_s| + o_K(1). \qquad (84)
$$

Scalar overlap q_t is defined similarly but replacing in (84) the matrices $(\boldsymbol{s}, \boldsymbol{S}, \boldsymbol{Q}_s)$ by $(\boldsymbol{t}, \boldsymbol{T}, \boldsymbol{Q}_t)$. Here the overlaps

$$
\boldsymbol{Q}_s := N^{-1}\boldsymbol{s}^\dagger\tilde{\boldsymbol{s}} \quad \text{and} \quad \boldsymbol{Q}_t := N^{-1}\boldsymbol{t}^\dagger\tilde{\boldsymbol{t}}
$$

for two i.i.d. samples $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{t})$ and $\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{t}})$ from the joint posterior distribution $P_{X|Y,N}$ given by (80) (these should not be confused with the samples drawn from the priors appearing in the statement below).

Result 8 (Replica symmetric formula for dictionary learning). Let $\tilde{p}_{S,K}$ by the j.p.d.f. of the K singular values [which are

$O(1)$] of $N^{-1}\boldsymbol{s}^\dagger\tilde{\boldsymbol{s}}$, where $\boldsymbol{s}, \tilde{\boldsymbol{s}}$ are i.i.d. $N \times K$ random matrices drawn from prior $P_{S,K}$. Similarly, let $\tilde{p}_{T,K}$ be the j.p.d.f. of the K singular values of $N^{-1}\boldsymbol{t}^\dagger\tilde{\boldsymbol{t}}$, where $\boldsymbol{t}, \tilde{\boldsymbol{t}}$ are i.i.d. $M \times K$ matrices drawn from $P_{T,K}$. The mutual information of model (78) verifies

$$
\frac{1}{MN} I\left[(\boldsymbol{S}, \boldsymbol{T}); \sqrt{\frac{\lambda}{N}}\boldsymbol{S}\boldsymbol{T}^\dagger + \boldsymbol{Z} \right]
$$

$$
= -\frac{K^2}{2MN} \sup_{(\sigma_s, \sigma_t)\in\mathcal{S}_K} \left[\frac{\ln \tilde{p}_{S,K}(\boldsymbol{\sigma}_s)}{K^2} + \frac{\ln \tilde{p}_{T,K}(\boldsymbol{\sigma}_t)}{K^2} \right.
$$

$$
\left. + J_K^{(\beta)}(\boldsymbol{\sigma}_s, \boldsymbol{\sigma}_t, 2\lambda) \right] + \frac{\beta\lambda}{2MN^2}\mathbb{E}\mathrm{Tr}\boldsymbol{S}^\dagger\boldsymbol{S}\boldsymbol{T}^\dagger\boldsymbol{T} + \tau_K. \quad (85)
$$

The set $\mathcal{S}_K$ is defined as

$$
\mathcal{S}_K := \big\{ (\boldsymbol{\sigma}_s, \boldsymbol{\sigma}_t) \in \mathbb{R}_{\geqslant 0}^K \times \mathbb{R}_{\geqslant 0}^K :
$$

$$
\boldsymbol{\sigma}_s \text{ is rank } \min(N, K), \ \boldsymbol{\sigma}_t \text{ is rank } \min(M, K)\big\}.
$$

Constant τ_K fixes $I((\boldsymbol{S}, \boldsymbol{T}); \boldsymbol{Z}) = 0$, i.e.,

$$\tau_K := \tfrac{1}{2} \sup_{(\boldsymbol{\sigma}_s, \boldsymbol{\sigma}_t) \in \mathcal{S}_K} g_K^{\mathrm{RS}}(\boldsymbol{\sigma}_s, \boldsymbol{\sigma}_t, \lambda = 0) + o_K(1),$$

where $g_K^{\mathrm{RS}} : \mathbb{R}_{\geqslant 0}^K \times \mathbb{R}_{\geqslant 0}^K \times \mathbb{R}_{\geqslant 0} \mapsto \mathbb{R}$ is the replica symmetric potential function defined by the curly brackets $\{\cdots\}$ in the variational problem (85). Denote $(\boldsymbol{\sigma}_s^*, \boldsymbol{\sigma}_t^*)$ the overlaps singular values achieving the supremum in (85). The scalar overlaps are

$$q_s = \frac{1}{N} \mathrm{Tr}\boldsymbol{\sigma}_s^* + o_K(1), \quad q_t = \frac{1}{N} \mathrm{Tr}\boldsymbol{\sigma}_t^* + o_K(1). \quad (86)$$

We can express the result using density order parameters, under an even weaker decoupling assumption than (83). The steps are similar to what is done in Sec. V E so we will be briefer. Let the empirical distributions of singular values

$$\hat{\rho}_{s,K}^{ab}(x) := \sum_{k \leqslant K} \frac{\delta(\sigma_{s,i}^{ab} - x)}{K}, \quad \hat{\rho}_{t,K}^{ab}(x) := \sum_{k \leqslant K} \frac{\delta(\sigma_{t,i}^{ab} - x)}{K}.$$

We assume that the random $(\hat{\rho}_{s,K}^{ab})_{a<b}$ jointly verify a LDP at scale K^2:

$$D\hat{P}_{S,K}\big[(\hat{\rho}_s^{ab})_{a<b}\big] \sim \exp\big(-K^2 \tilde{L}_s\big[(\hat{\rho}_s^{ab})_{a<b}\big]\big) \prod_{a<b}^{0,u} D\big[\hat{\rho}_s^{ab}\big],$$

and similarly for $(\hat{\rho}_{t,K}^{ab})_{a<b}$ with rate functional $\tilde{L}_t$. Then the weaker decoupling assumption reads

Assumption (spectral replica symmetry, weaker form):

$$\tilde{L}_{s/t}\big[(\hat{\rho}_{s/t}^{ab})_{a<b}\big] \to \sum_{a<b}^{0,u} L_{s/t}\big[\hat{\rho}_{s/t}^{ab}\big].$$

The rate functional L_s is naturally taken to be the one corresponding to the empirical distribution of singular values of a product of two i.i.d. samples from the prior $P_{S,K}$, and similarly for L_t but with samples from $P_{T,K}$. Start again from (82). Absorb the singular vectors of the overlaps in the Haar distributed matrices $\boldsymbol{U}^{ab}$ and $\boldsymbol{V}^{ab}$, so that the rectangular spherical integrals appearing in (82) only depend on the singular values $(\boldsymbol{\sigma}_s^{ab}, \boldsymbol{\sigma}_t^{ab})_{a<b}$. Only the measures $dP_{(\sigma_s),K}[(\sigma_s^{ab})_{a<b}] dP_{(\sigma_t),K}[(\sigma_t^{ab})_{a<b}]$ then remain. These induce j.p.d.f. over the above empirical distributions of singular values. As seen from (50), asymptotically, the rectangular spherical depends only on the density of singular values. All these observations combined with the aforementioned LDPs and decoupling ansatz (VI B) allow us to simplify (82) into

$$\mathbb{E}\mathcal{Z}^u \propto \left[\int D[\hat{\rho}_s] D[\hat{\rho}_t] \exp K^2 (-L_s[\hat{\rho}_s] - L_t[\hat{\rho}_t] \right.$$
$$\left. + J^{(\beta)}[\hat{\rho}_s, \hat{\rho}_t, 2\lambda] + o_K(1)) \right]^{u(u+1)/2}.$$

Saddle-point estimation and (55) then yields the following variational formulation.

Result 9 (Replica symmetric formula for dictionary learning). Assume that the empirical distribution of singular values of $N^{-1}\boldsymbol{s}^\dagger \tilde{\boldsymbol{s}}$, where $\boldsymbol{s}, \tilde{\boldsymbol{s}}$ are i.i.d. $N \times K$ random matrices drawn from prior $P_{S,K}$, verifies a LDP at scale K^2 with rate functional L_s. Similarly, assume that the empirical distribution of

singular values of $N^{-1}\boldsymbol{t}^\dagger \tilde{\boldsymbol{t}}$, where $\boldsymbol{t}, \tilde{\boldsymbol{t}}$ are i.i.d. $M \times K$ from prior $P_{S,K}$, verifies a LDP at scale K^2 with rate functional L_t. The mutual information verifies, in the limit $K \to +\infty$ with $N/K \to \alpha$ and $M/K \to \gamma$,

$$\frac{1}{MN} I\left[(\boldsymbol{S}, \boldsymbol{T}); \sqrt{\frac{\lambda}{N}} \boldsymbol{S}\boldsymbol{T}^\dagger + \boldsymbol{Z} \right]$$

$$\to -\frac{1}{2\alpha\gamma} \sup_{(\rho_s, \rho_t) \in \mathcal{S}} [J^{(\beta)}[\rho_s, \rho_t, 2\lambda] - L_s[\rho_s] - L_t[\rho_t]]$$

$$+ \frac{\beta\lambda}{2} \lim_{K \to +\infty} \frac{\mathbb{E}\mathrm{Tr}\boldsymbol{S}^\dagger \boldsymbol{S}\boldsymbol{T}^\dagger \boldsymbol{T}}{MN^2} + \tau. \quad (87)$$

The optimization is over probability densities with finite non-negative support belonging to

$$\mathcal{S} := \{(\rho_s, \rho_t) = \{[1 - \min(1, \alpha)]\delta_0 + \min(1, \alpha)\tilde{\rho}_s,$$
$$[1 - \min(1, \gamma)]\delta_0 + \min(1, \gamma)\tilde{\rho}_t\}, (\tilde{\rho}_s, \tilde{\rho}_t) \in \mathcal{P}_{\geqslant 0} \times \mathcal{P}_{\geqslant 0}\}.$$

Constant τ fixes the constraint $I((\boldsymbol{S}, \boldsymbol{T}); \boldsymbol{Z}) = 0$, i.e.,

$$\tau := \frac{1}{2\alpha\gamma} \sup_{(\rho_s, \rho_t) \in \mathcal{S}} g^{\mathrm{RS}}[\rho_s, \rho_t, \lambda = 0],$$

for the replica symmetric potential functional $g^{\mathrm{RS}} : \mathcal{P}_{\geqslant 0} \times \mathcal{P}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0} \mapsto \mathbb{R}$ defined by the curly brackets in (87). The densities achieving the supremum in (87) are

$$(\rho_s^*, \rho_t^*) = [(1 - \min(1, \alpha))\delta_0 + \min(1, \alpha)\tilde{\rho}_s^*,$$
$$[1 - \min(1, \gamma)]\delta_0 + \min(1, \gamma)\tilde{\rho}_t^*).$$

The overlaps are then given by

$$q_s = \int d\rho_s^*(x) x, \quad q_t = \int d\rho_t^*(x) x.$$

Note that in the more symmetric special case of $M = N$ and $P_{S,K} = P_{T,K}$ [which does *not* correspond to the Hermitian dictionary learning problem (52) as $\boldsymbol{S}$ and $\boldsymbol{T}$ remain i.i.d.], by symmetry, the replica symmetric formula can be simplified as (setting $L_s = L_t = L$)

$$\frac{1}{MN} I\left[(\boldsymbol{S}, \boldsymbol{T}); \sqrt{\frac{\lambda}{N}} \boldsymbol{S}\boldsymbol{T}^\dagger + \boldsymbol{Z} \right]$$

$$\to -\sup_{\rho \in \mathcal{S}_{\mathrm{diag}}} \left[\frac{J^{(\beta)}[\rho, \rho, 2\lambda]}{2\alpha^2} - \frac{L[\rho]}{\alpha^2} \right]$$

$$+ \frac{\beta\lambda}{2} \lim_{K \to +\infty} \frac{\mathbb{E}\mathrm{Tr}\boldsymbol{S}^\dagger \boldsymbol{S}\boldsymbol{T}^\dagger \boldsymbol{T}}{N^3} + \tau_{\mathrm{diag}},$$

where $\mathcal{S}_{\mathrm{diag}}$ is the "diagonal subset" of $\mathcal{S}$ where additionally $\rho_s = \rho_t$ (and the definition of τ_{diag} is modified from τ accordingly). The unique scalar overlap is in this case $q_s = q_t = \int d\rho^*(x) x$ where ρ^* achieves the supremum in the variational problem for the mutual information.

VII. COMPARISON TO PREVIOUS WORKS, CONCLUSION AND OPEN PROBLEMS

The Bayes-optimal setting of linear-rank dictionary learning has been previously studied in the inspiring works [52–56] (in the real case $\beta = 1$). But these approaches provide only approximations to the exact asymptotic formulas. In [55] they

assume that the entries of $\boldsymbol{ST}^\intercal$ in (78) are jointly Gaussian while they are not; see [57] for more explanations. It also corresponds to assume that the overlap matrix $\boldsymbol{Q}^{ab}$ is parametrized by a single number. In his thesis [56] Schmidt went further. He proposed instead that the overlap really is a matrix but symmetric (while there is no reason for it to be so) and completely independent of the replica indices. This assumption that both eigenvalues and eigenvectors are replica independent is physically equivalent to assume that the overlaps $(\boldsymbol{Q}^{ab})_{a<b}$ concentrate entrywise. However, we expect that only the statistics of eigenvalues or singular values can be self-averaging as is often the case in random matrix theory. The same phenomenon happens in large covariances matrices: only the spectral properties become deterministic while the matrix entries fluctuate even in the large size limit.

In the most generic version of the method leading to Result 5, our ansatz is only at the level of the distribution of singular values of the overlaps $\boldsymbol{Q}^{ab} = \boldsymbol{A}^{ab}\boldsymbol{\sigma}_Q^{ab}\boldsymbol{B}^{ab}$: the singular values of the matrices $\boldsymbol{\sigma}_Q^{ab}$ are assumed to decouple and to have identical statistics in the large size limit. Nothing is assumed about the singular vectors. The spectral decoupling assumption (61) allows us to carry on the computation while completely capturing the relevant rotational degrees of freedom and invariances. The spectral replica method therefore allows us to reduce the challenging task of computing the quenched free entropy (or mutual information), which is an integral over $\Theta(N^2)$ matrix elements (and that additionally should be averaged over the data distribution), into two well defined RMT subproblems:

(1) Obtaining the j.p.d.f. of the singular values (or eigenvalues) of a product of two i.i.d. random matrices drawn from the prior distribution $P_{X,N}$. Or alternatively, to obtain the rate functional controlling the large deviations of the empirical spectral distribution.

(2) Analyzing a Coulomb gas, i.e., an optimization problem over $\Theta(N)$ interacting degrees of freedom representing the singular values (or eigenvalues) of certain matrix order parameters entering the analysis. Or equivalently, solving a functional optimization problem over the associated asymptotic densities.

As already discussed, the first task does not require *a priori* the prior $P_{X,N}$ to be factorized over the entries of the matrix $\boldsymbol{X}$. So in cases where the j.p.d.f. p_M (or the rate functional L) can be evaluated, the spectral replica method yields concrete asympotic formulas for the mutual information and MMSE. To conclude, we have studied two classes of matrix inference problems in the Bayesian-optimal setting: *matrix denoising*, and symmetric and nonsymmetric *dictionary learning*. Most importantly, they are analyzed in the challenging scaling regime of *linear-rank* matrix signals for which very few prior results exist. These problems bear some similarity with matrix models originating in high-energy physics, but with one crucial aspect when applied to inference, namely, the presence of a quenched average over the data. The analysis of matrix denoising could be carried out using known tools from random matrix theory, in particular the use of spherical integrals and related perturbative expansions. On the other hand for dictionary learning our analysis combines the use of spherical integrals with a replica symmetric ansatz, which we call the *spectral replica method*. For dictionary learning, the

solutions are expressed in terms of nontrivial variational problems reminiscent of Coulomb gases (but with extra complications), which reduce the number of degrees of freedom from $\Theta(N^2)$ down to $\Theta(N)$. Finding numerical solutions of the variational problems appears to be highly nontrivial. There are two difficulties: one due to the appearance of spherical integrals, and a second due to the determination of the j.p.d.f. of products of i.i.d. random matrices (the distribution p_M), or of the rate functional associated with the large deviations of their empirical spectral densities. We have outlined in Sec. V F how the second problem can sometimes be reduced to the calculation of other spherical integrals. We conjecture that our solutions of the denoising and dictionary learning problems are exact in the thermodynamic limit. In Sec. V D we have presented a sanity check in a simple case. However this case hardly constitutes a satisfying check. Finding other more interesting checks remains one of the most important problems. A more promising and much less trivial check has been outlined in Sec. V F. The proposed strategy is based on the method of biorthogonal polynomials for exactly computing the integral resulting from our analysis in the case of Hermitian dictionary learning with a Ginibre signal, which will then be comparable against our predictions for matrix denoising using Matytsin's hydrodynamic mapping [88]. While we believe that this is not out of reach, it is certainly not an easy task. The variational problems can also form the starting point of perturbative expansions in regimes of small and large signal-to-noise ratio, and/or expansions in powers of the inverse rank. It may then be possible to obtain nontrivial comparisons with the "high-temperature" expansions derived in [57] based on the Plefka-George-Yedidia formalism. As a matter of fact, even if our proposed solutions were not exact, since they are based on a rather weak form of replica symmetric ansatz (the spectral replica symmetry) they could still constitute a good approximation, that would be useful in nonperturbative regimes. Finally, we hope that eventually our study can inspire rigorous progress on these problems. A first desirable step would be to provide proofs of Conjectures 3 and 4 for the easier denoising problem.

APPENDIX A: SPHERICAL INTEGRALS

In this Appendix we present the HCIZ formula in the Hermitian and general non-Hermitian matrix cases. For the derivations we refer to the very readable original papers by Itzykson and Zuber [63] and Mehta [66].

1. Hermitian case

Consider Hermitian $M \times M$ matrices $\boldsymbol{A}, \boldsymbol{B}$. These are diagonalized by unitary matrices and have real eigenvalues. Recall the notation $\boldsymbol{A} = \boldsymbol{U}^A\boldsymbol{\lambda}^A(\boldsymbol{U}^A)^\dagger$, $\boldsymbol{B} = \boldsymbol{U}^B\boldsymbol{\lambda}^B(\boldsymbol{U}^B)^\dagger$. Recall the

definition of the Vandermonde (5). The HCIZ formula reads

$$\int_{\mathcal{U}(M)} d\mu_M^{(2)}(U) \exp \gamma M \mathrm{Tr}(A U^\dagger B U)$$

$$= \frac{\prod_{k \leqslant M-1} k!}{(\gamma M)^{M(M-1)/2}} \frac{\det(\exp \gamma M \lambda_i^A \lambda_j^B)}{\Delta_M(\lambda^A) \Delta_M(\lambda^B)}, \qquad (A1)$$

where $\mu_M^{(2)}$ is the normalized Haar measure over the group of unitary $M \times M$ matrices and $[\exp \gamma M \lambda_i^A \lambda_j^B]$ is the matrix $[\exp(\gamma M \lambda_i^A \lambda_j^B)]_{i,j \leqslant M}$. Note that on the left-hand side we can replace A, B by λ^A, λ^B since U_A, U_B leave the Haar measure invariant. Note also that by permutation symmetry the ratio of determinants is positive and independent of the ordering of eigenvalues. In the limit $M \to +\infty$ the spherical integral can be described in terms of an hydrodynamical system (the complex Burgers equation) thanks to the work of Matytsin [88] and proven in [87]. See also [102,130]. This is not used in this paper but may be useful for future analyses.

2. General non-Hermitian case

Let A, B be two general $M \times M$ matrices. Their singular value decomposition is $A = U^A \sigma^A V^A$ and $B = U^B \sigma^B V^B$ where σ^A, σ^B are the diagonal matrices of nonnegative singular values and U^A, V^A, U^B, and V^B are unitary matrices. The spherical integral involves the modified Bessel function of the first kind:

$$I_0(x) := \int_0^\pi \frac{d\theta}{\pi} \exp(x \cos \theta).$$

This function is positive, monotone increasing, $I_0(0) = 1$, and grows as $\exp x$ at infinity. We have

$$\int_{\mathcal{U}(M) \times \mathcal{U}(M)} d\mu_M^{(2)}(U) \, d\mu_M^{(2)}(V) \exp \gamma M \mathrm{Re}\mathrm{Tr}(A U B V)$$

$$= \frac{2^{M(M-1)} \left(\prod_{k \leqslant M-1} k!\right)^2}{M! (M\gamma)^{M(M-1)}} \frac{\det\left[I_0\left(M\gamma \sigma_i^A \sigma_j^B\right)\right]}{\Delta_M[(\sigma^A)^2] \Delta_M[(\sigma^B)^2]}.$$

As before, on the left-hand side we can replace A, B by σ^A, σ^B, and the ratio of determinants is positive and invariant under permutations of singular values. This formula first obtained in [108] and proven in [109] can be derived by the same methods used for the classical HCIZ formula based on the solution of the heat equation. Like for the standard spherical integral, there also exists an asymptotic $M \to +\infty$ representation of the rectangular spherical integral in terms of a complex hydrodynamical system [110,111].

APPENDIX B: MINIMUM MEAN-SQUARE ERROR: A MORE EXPLICIT FORMULA

By the I-MMSE relation (12), we obtained in Result 3 that the MMSE is directly proportional to the derivative with respect to the signal-to-noise ratio λ of the HCIZ integral. Its dependence in λ is through the data-matrix eigenvalues λ^Y. From the HCIZ formula (see Appendix A) we get in the complex case $\beta = 2$, using the Jacobi formula

$$\frac{\partial}{\partial X_{ij}} \ln \det X = (X^{-1})_{ji},$$

that the HCIZ derivative verifies

$$N^2 \frac{d}{d\lambda} I_N^{(2)}(\lambda^S, \lambda^Y, \sqrt{\lambda}) = \frac{N^2}{2\sqrt{\lambda}} \frac{d}{d\sqrt{\lambda}} I_N^{(2)}(\lambda, \sqrt{\lambda}\lambda^Y, 1)$$

$$= \frac{1}{2\sqrt{\lambda}} \frac{d}{d\sqrt{\lambda}} \ln \det\left[\exp N \lambda_c^S\left(\sqrt{\lambda}\lambda_d^Y\right)\right] - \frac{1}{2\sqrt{\lambda}} \frac{d}{d\sqrt{\lambda}} \ln \det\left[\left(\sqrt{\lambda}\lambda_c^Y\right)^{d-1}\right]$$

$$= \frac{1}{2\sqrt{\lambda}} \sum_{i,j \leqslant N} \frac{d \ln \det\left[\exp N \lambda_c^S\left(\sqrt{\lambda}\lambda_d^Y\right)\right]}{d \exp N \lambda_i^S\left(\sqrt{\lambda}\lambda_j^Y\right)} \frac{d \exp N \lambda_i^S\left(\sqrt{\lambda}\lambda_j^Y\right)}{d\sqrt{\lambda}}$$

$$- \frac{1}{2\sqrt{\lambda}} \sum_{i,j \leqslant N} \frac{d \ln \det\left[\left(\sqrt{\lambda}\lambda_c^Y\right)^{d-1}\right]}{d\left(\sqrt{\lambda}\lambda_i^Y\right)^{j-1}} \frac{d\left(\sqrt{\lambda}\lambda_i^Y\right)^{j-1}}{d\sqrt{\lambda}}$$

$$= \frac{1}{2\sqrt{\lambda}} \sum_{i,j \leqslant N} \left\{\left[\exp N \lambda_c^S\left(\sqrt{\lambda}\lambda_d^Y\right)\right]^{-1}\right\}_{ji} N \lambda_i^S \left[\exp N \lambda_i^S\left(\sqrt{\lambda}\lambda_j^Y\right)\right] \frac{d\sqrt{\lambda}\lambda_j^Y}{d\sqrt{\lambda}}$$

$$- \frac{1}{2\sqrt{\lambda}} \sum_{i,j \leqslant N} \left\{\left[\left(\sqrt{\lambda}\lambda_c^Y\right)^{d-1}\right]^{-1}\right\}_{ji} (j-1)\left(\sqrt{\lambda}\lambda_i^Y\right)^{j-2} \frac{d\sqrt{\lambda}\lambda_i^Y}{d\sqrt{\lambda}}.$$

Introducing the Y eigenvectors $Y\psi_i^Y = \lambda_i^Y \psi_i^Y$, the Hellmann-Feynman theorem implies

$$\frac{d\lambda_i^Y}{d\sqrt{\lambda}} = \left(\psi_i^Y\right)^\dagger S \psi_i^Y.$$

Thus

$$\frac{d\sqrt{\lambda}\lambda_i^Y}{d\sqrt{\lambda}} = \left(\psi_i^Y\right)^\dagger (\sqrt{\lambda}S + \xi)\psi_i^Y + \sqrt{\lambda}\left(\psi_i^Y\right)^\dagger S \psi_i^Y$$

$$= \left(\psi_i^Y\right)^\dagger (2\sqrt{\lambda}S + \xi)\psi_i^Y.$$

Putting everything together in (25) we find that the MMSE equals (when $\beta = 2$)

$$\frac{1}{N^2}\mathbb{E}\|S - \mathbb{E}[S|Y]\|^2 = \frac{4}{N}\mathbb{E}\mathrm{Tr}\lambda_S^2 - \frac{2}{\sqrt{\lambda}N^2}\sum_{i,j\leqslant N}\left\{\left[\exp N\lambda_c^S(\sqrt{\lambda}\lambda_d^Y)\right]^{-1}\right\}_{ji}N\lambda_i^S\left[\exp N\lambda_i^S(\sqrt{\lambda}\lambda_j^Y)\right](\boldsymbol{\psi}_j^Y)^\dagger(2\sqrt{\lambda}S + \boldsymbol{\xi})\boldsymbol{\psi}_j^Y$$

$$+ \frac{2}{\sqrt{\lambda}N^2}\sum_{i,j\leqslant N}\left\{\left[(\sqrt{\lambda}\lambda_c^Y)^{d-1}\right]^{-1}\right\}_{ji}(j-1)(\sqrt{\lambda}\lambda_i^Y)^{j-2}(\boldsymbol{\psi}_i^Y)^\dagger(2\sqrt{\lambda}S + \boldsymbol{\xi})\boldsymbol{\psi}_i^Y + o_N(1).$$

Besides the matrix inversions this formula also requires to compute eigenvectors of $Y = \sqrt{\lambda}S + \boldsymbol{\xi}$; it may be more practical to use $(\boldsymbol{\psi}_i^Y)^\dagger(2\sqrt{\lambda}S + \boldsymbol{\xi})\boldsymbol{\psi}_i^Y = \lambda_i^Y + \sqrt{\lambda}(\boldsymbol{\psi}_i^Y)^\dagger S\boldsymbol{\psi}_i^Y$. If one were to start instead from Result 1 the computation would be similar, because at the stationary point, the total λ derivative is computed by taking a partial derivative only with respect to the explicit λ dependence. Indeed, the terms coming from the implicit dependence in the solution $\boldsymbol{\lambda}^s$ of the fixed point equations do not contribute.

APPENDIX C: MATHEMATICA CODES

1. Small signal-to-noise expansion

The two first functions provided below allowing one to convert moments to free cumulants and vice versa are taken from [131]. This function gives the free cumulants as function of generic moments (M_i) of a density:

```
ln[1]:=m[z_] := 1 + M1z + M2z^2 + M3z^3
       +M4z^4 + M5z^5
       +M6z^6 + M7z^7 + M8z^8;
       Simplify[Table[{k, (−(k−1)^(−1)/k!)
       *D[m[z]^(1−k), z, k]/.z− > 0}, {k, 2, 8}]]
```

The next function gives the moment generating function expressed with the free cumulants and thus allows us to read the expression of the moments in terms of free cumulants. We force the values of the first two free cumulants to $k_1 = 0$ and $k_2 = 1$ (but this is not necessary):

```
lm[2] := k1 = 0; k2 = 1; m[z] = 1;
        r[z] := k1 + k2z + k3z^2 + k4z^3
        +k5z^4 + k6z^5
        +k7z^6 + k8z^7 + k9z^8;
        Do[rtmp[z_] = PolynomialMod[zr[z],
        z^(k + 1)];
        m[z_] = PolynomialMod[1 +
        rtmp[zm[z]], z^(k + 1)],
        k, 0, 8];
```

```
Collect[m[z], z]
```

We now express the asymptotic moments $m_i = \theta_i := \lim_{N\to+\infty}N^{-1}\mathrm{Tr}S^i$ of the signal S as a function of the free cumulants (k_i) of its asymptotic spectral density due to the previous function. We focus on traceless signals $m_1 = 0 = k_1$ and with normalized variance $m_2 = 1 = k_2$; the first condition does not change anything from the information-theoretic point of view as explained in Remark 1 below Result 3, and the second condition simply amounts to a rescaling of λ if not *a priori* verified:

```
In[3] := m3 = k3;
         m4 = k4 + 2;
         m5 = 5k3 + k5;
         m6 = 3k3^2 + 6k4 + k6 + 5;
         m7 = 7k3k4 + 21k3 + 7k5 + k7;
         m8 = 28k3^2 + 8k3k5 + 4k4^2
         +28k4 + 8k6 + k8 + 14;
```

We express the free cumulants (c_i) of the data Y as a function of the free cumulants (k_i) of the signal S. The (c_i) are the $(\bar{\phi}_i)$ in the Zinn-Justin and Zuber expansion [106] (see also [105] for the same expansion in terms of trace moments, also found in [106]). The Wigner matrix $\boldsymbol{\xi}$ only shifts the second free cumulant of $\sqrt{\lambda}S$ (which is λ) by 1. Below, variable snr refers to λ:

```
In[4] := c2 = snr + 1;
         c3 = k3snr^(3/2);
         c4 = k4snr^2;
         c5 = k5snr^(5/2);
         c6 = k6snr^3;
         c7 = k7snr^(7/2);
         c8 = k8snr^4;
```

The terms (F_n) in the Zinn-Justin and Zuber expansion of the spherical integral [106], expressed with the free cumulants (c_i) of the data matrix, and the moments (m_i) of the signal S; the (m_i) are the (θ_i) in the expansion found in [106]:

```
In[5]F2 = c2/2;
        F3 = c3m3/3;
        F4 = c4m4/4 − 1/2(c2^2/2 + c4);
        F5 = c5m5/5 − m3(c2c3 + c5);
```

$$F6 = -1/2 m3^2(c2^3/3 + c2c4 + c3^2 + c6) + 1/6(2c2^3 + 12c2c4 + 5c3^2 + 7c6)$$
$$-m4(c2c4 + c3^2/2 + c6) + (c6m6)/6;$$
$$F7 = -m3m4(c2^2c3 + c2c5 + 2c3c4 + c7) + m3(5c2^2c3 + 7c2c5 + 8c3c4 + 4c7)$$
$$-m5(c2c5 + c3c4 + c7) + c7m7/7;$$
$$F8 = -m3m5(c2^2c4 + c2c3^2 + c2c6 + 2c3c5 + c4^2 + c8) + m3^2(2c2^4$$
$$+16c2^2c4 + 20c2c3^2 + 16c2c6 + 24c3c5 + 11c4^2 + 9c8)$$
$$-1/2 m4^2(c2^4/4 + c2^2c4 + 2c2c3^2 + c2c6 + 2c3c5 + 3/2c4^2 + c8)$$
$$+1/2 m4(c2^4 + 11c2^2c4 + 14c2c3^2 + 16c2c6 + 18c3c5 + 11c4^2 + 9c8)$$
$$-3/8(3c2^4 + 24c2^2c4 + 24c2c3^2 + 24c2c6 + 24c3c5 + 15c4^2 + 10c8)$$
$$-m6(c2c6 + c3c5 + c4^2/2 + c8) + c8m8/8;$$

The expansion of the mutual information is, according to our Result 3, given up to $O(\lambda^4)$ by

```
In[6] := MI8 = snr − F2snr − F3snr^(3/2)
         −F4snr^2 − F5snr^(5/2)
         −F6snr^3 − F7snr^(7/2)
         −F8snrS^4;
         MutInfo = Collect[Simplify[MI8], snr];
```

Only the first four orders are reliable. This code gives the generic expansions (46) and (47) in the case $m_1 = 0$ and $m_2 = 1$ (but this can be easily adapted using the code).

2. Useful code to produce Figs. 1 and 2

This parts evaluates the spectral density of the data matrix **Y** by solving the transcendental equation (34) for its Green function. The spectral density is then extracted from its imaginary part:

```
In[7] := snr = 1; step = 0.0005; zAndrho = {};
         init = I/5; bound = 4;
         Do[zAndrho = Append[zAndrho,
         {z, Abs[Im[g/.FindRoot[SetAccuracy[z == Sqrt[3snr]
         *Coth[gSqrt[3snr]] + g, 30], g, init, WorkingPrecision− > 20]]]/Pi}],
         {z, −bound, bound, step}];
```

As there may be multiple solutions depending on the initial point `init` for the search (that may need to be tuned), a sanity check is to check that the solution found is properly normalized:

```
In[8] := Print["Normalization = ",
         Total[zAndrho[[All, 2]]]step];
```

Now we find an interpolating function for the spectral density of **Y** from the previously equally spaced computed points, using Hermite polynomials. Plotting this interpolating function is what gives the asymptotic red curves in Fig. 1:

```
In[9] := z = zAndrho[[All, 1]];
         rho = Chop[zAndrho[[All, 2]]];
         rhoInterp = ListInterpolation[rho,
         Min[z], Max[z], Method− > "Hermite"];
```

We can now compute the asymptotic mutual information using (33), based on the interpolation function, and compare the results to the Wigner case:

```
In[10] := MI = snr + Log[snr12]/4
          +0.5NIntegrate[rhoInterp[x]
          *rhoInterp[y]Log[Abs[(x − y)
          /(Exp[xSqrt[12snr]]
          −Exp[ySqrt[12snr]])]],
          {x, Min[z], Max[z]},
          {y, Min[z], Max[z]}];
          Print[Abs[MI − 0.5Log[1 + snr]]];
```

Using these pieces of code and running them for various λ, one can obtain the pink dots in Fig. 2. The finite-size curves (blue and orange dots) are instead simply obtained by averaging the associated formulas over many large realizations of the model.

[1] T. Tao, *Topics in Random Matrix Theory*, Graduate Studies in Mathematics, Vol. 132 (American Mathematical Society, Providence, RI, 2012), p. 282.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge Studies in Advanced Mathematics, Vol. 118 (Cambridge University Press, Cambridge, 2010).

[3] M. L. Mehta, *Random Matrices*, Pure and Applied Mathematics, Vol. 142 (Academic Press, 2004).

[4] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory* (Cambridge University Press, Cambridge, 2021).

[5] G. Livan, M. Novaes, and P. Vivo, *Introduction to Random Matrices: Theory and Practice*, SpringerBriefs in Mathematical Physics, Vol. 26 (Springer, 2018).

[6] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, Online dictionary learning for sparse coding, in *Proceedings of the 26th Annual International Conference on Machine Learning* (Association for Computing Machinery, New York, 2009), pp. 689–696.

[7] I. Tošić and P. Frossard, Dictionary learning, IEEE Signal Proc. Mag. **28**, 27 (2011).

[8] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).

[9] B. A. Olshausen and D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature (London) **381**, 607 (1996).

[10] B. A. Olshausen and D. J. Field, Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision Res. **37**, 3311 (1997).

[11] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, Dictionary learning algorithms for sparse representation, Neural Comput. **15**, 349 (2003).

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, Robust principal component analysis? J. ACM **58**, 1 (2011).

[13] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, Optimality and sub-optimality of PCA I: Spiked random matrix models, Ann. Stat. **46**, 2416 (2018).

[14] B. Hajek, Y. Wu, and J. Xu, Information limits for recovering a hidden community, IEEE Trans. Inf. Theory **63**, 4729 (2017).

[15] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, A blind source separation technique using second-order statistics, IEEE Trans. Signal Process. **45**, 434 (1997).

[16] E. Candès and B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. **9**, 717 (2009).

[17] E. J. Candès and T. Tao, The power of convex relaxation: Near-optimal matrix completion, IEEE Trans. Inf. Theory **56**, 2053 (2010).

[18] E. Abbe, Community detection and stochastic block models: Recent developments, J. Machine Learn. Res. **18**, 1 (2018).

[19] M. Lelarge and L. Miolane, Fundamental limits of symmetric low-rank matrix estimation, Prob. Theory Related Fields **173**, 859 (2019).

[20] F. Caltagirone, M. Lelarge, and L. Miolane, Recovering asymmetric communities in the stochastic block model, IEEE Trans. Netw. Sci. Eng. **5**, 237 (2017).

[21] I. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. **29**, 295 (2001).

[22] I. Johnstone and A. Lu, Sparse principal components analysis, arXiv:0901.4392 [math.ST].

[23] H. Zou, T. Hastie, and R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stati. **15**, 265 (2006).

[24] I. Johnstone and A. Lu, On consistency and sparsity for principal components analysis in high dimensions, J. Am. Stat. Assoc. **104**, 682 (2009).

[25] J. Baik, G. B. Arous, and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, Ann. Probab. **33**, 1643 (2005).

[26] J. Baik and J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, J. Multivariate Anal. **97**, 1382 (2006).

[27] S. B. Korada and N. Macris, Exact solution of the gauge symmetric *p*-spin glass model on a complete graph, J. Stat. Phys. **136**, 205 (2009).

[28] T. Lesieur, F. Krzakala, and L. Zdeborová, Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications, J. Stat. Mech.: Theory Exp. (2017) 073403.

[29] Y. Deshpande, E. Abbe, and A. Montanari, Asymptotic mutual information for the balanced binary stochastic block model, Information Inference **6**, 125 (2017).

[30] L. Miolane, Fundamental limits of low-rank matrix estimation: The non-symmetric case, arXiv:1702.00473 (2017).

[31] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, Statistical and computational phase transitions in spiked tensor estimation, in *IEEE International Symposium on Information Theory (ISIT)* (IEEE, New York, 2017), pp. 511–515.

[32] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula, in *Advances in Neural Information Processing Systems (NIPS)*, Vol. 29 (Association for Computing Machinery, New York, 2016), pp. 424–432.

[33] J. Barbier and N. Macris, The adaptive interpolation method: A simple scheme to prove replica formulas in Bayesian inference, Probab. Theory Relat. Fields **174**, 1133 (2019).

[34] J. Barbier and N. Macris, The adaptive interpolation method for proving replica formulas. Applications to the Curie–Weiss and Wigner spike models, J. Phys. A: Math. Theor. **52**, 294002 (2019).

[35] J. Barbier, M. Dia, N. Macris, F. Krzakala, and L. Zdeborová, Rank-one matrix estimation: Analysis of algorithmic and information theoretic limits by the spatial coupling method, arXiv:1812.02537 (2018).

[36] J. Barbier, N. Macris, and L. Miolane, The layered structure of tensor estimation and its mutual information, in 47th Annual Allerton Conference on Communication, Control, and Computing, arXiv:1709.10368 (2017)

[37] A. E. Alaoui, F. Krzakala, and M. I. Jordan, Finite size corrections and likelihood ratio fluctuations in the spiked Wigner model, arXiv:1710.02903 (2017).

[38] A. El Alaoui and F. Krzakala, Estimation in the spiked Wigner model: A short proof of the replica formula, in *2018 IEEE International Symposium on Information Theory (ISIT)* (IEEE, New York, 2018), pp. 1874–1878.

[39] J.-C. Mourrat, Hamilton-Jacobi equations for mean-field disordered systems, arXiv:1811.01432 (2018).

[40] J.-C. Mourrat, Hamilton-Jacobi equations for finite-rank matrix inference, arXiv:1904.05294 (2019).

[41] J. Barbier, N. Macris, and C. Rush, All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates Inc., Red Hook, NY, USA, 2020).

[42] G. Reeves, Information-theoretic limits for the matrix tensor product, IEEE J. Selected Areas Inf. Theory **1**, 777 (2020).

[43] D. Alberici, F. Camilli, P. Contucci, and E. Mingione, The multi-species mean-field spin-glass on the Nishimori line, J. Stat. Phys. **182**, 2 (2021).

[44] F. Camilli, P. Contucci, and E. Mingione, An inference problem in a mismatched setting as a spin-glass model with Mattis interaction, SciPost Phys. **12**, 125 (2022).

[45] F. Pourkamali and N. Macris, Mismatched estimation of rank-one symmetric matrices under Gaussian noise, arXiv:2107.08927 (2021).

[46] D. Alberici, F. Camilli, P. Contucci, and E. Mingione, The solution of the deep Boltzmann machine on the Nishimori line, Commun. Math. Phys. **387**, 1191 (2021).

[47] S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, and L. Zdeborová, Who is afraid of big bad minima? Analysis of gradient-flow in spiked matrix-tensor models, in *Advances in Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019), pp. 8676–8686.

[48] S. S. Mannelli, F. Krzakala, P. Urbani, and L. Zdeborova, Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models, in *International Conference on Machine Learning* (PMLR, Cambridge, MA, 2019), pp. 4333–4342.

[49] T. Liu, Y. Li, S. Wei, E. Zhou, and T. Zhao, Noisy gradient descent converges to flat minima for nonconvex matrix factorization, arXiv:2102.12430 (2021).

[50] A. Maillard, L. Foini, A. L. Castellanos, F. Krzakala, M. Mézard, and L. Zdeborová, High-temperature expansions and message passing algorithms, J. Stat. Mech.: Theory Exp. (2019) 113301.

[51] A. Bodin and N. Macris, Rank-one matrix estimation: analytic time evolution of gradient descent dynamics, arXiv:2105.12257 (2021).

[52] A. Sakata and Y. Kabashima, Statistical mechanics of dictionary learning, Europhys. Lett. **103**, 28008 (2013).

[53] A. Sakata and Y. Kabashima, Sample complexity of Bayesian optimal dictionary learning, in *2013 IEEE International Symposium on Information Theory* (IEEE, Istambul, Turkey, 2013), p. 669.

[54] F. Krzakala, M. Mézard, and L. Zdeborová, Phase diagram and approximate message passing for blind calibration and dictionary learning, in *2013 IEEE International Symposium on Information Theory* (IEEE, Istambul, Turkey, 2013), p. 659.

[55] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, Phase transitions and sample complexity in Bayes-optimal matrix factorization, IEEE Trans. Inf. Theory **62**, 4228 (2016).

[56] H. C. Schmidt, Statistical physics of sparse and dense models in optimization and inference, Ph.D. thesis, Universit Paris Saclay (COmUE), 2018, https://tel.archives-ouvertes.fr/tel-03227132.

[57] A. Maillard, F. Krzakala, M. Mézard, and L. Zdeborová, Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising, J. Stat. Mech. (2022) 083301.

[58] J. T. Parker, P. Schniter, and V. Cevher, Bilinear generalized approximate message passing—Part I: Derivation, IEEE Trans. Signal Proc. **62**, 5839 (2014).

[59] J. T. Parker, P. Schniter, and V. Cevher, Bilinear generalized approximate message passing—Part II: Applications, IEEE Trans. Signal Proc. **62**, 5854 (2014).

[60] J. Bun, R. Allez, J.-P. Bouchaud, and M. Potters, Rotational invariant estimator for general noisy matrices, IEEE Trans. Inf. Theory **62**, 7475 (2016).

[61] It is now known that the estimator proposed in paper [60] is Bayes-optimal in certain settings; see [57].

[62] Harish-Chandra, Differential operators on a semisimple Lie algebra, Am. J. Math. **79**, 87 (1957).

[63] C. Itzykson and J.-B. Zuber, The planar approximation. II, J. Math. Phys. **21**, 411 (1980).

[64] P. Bleher and A. Its, *Random Matrix Models and Their Applications*, Mathematical Sciences Research Institute Publications, Vol. 40 (Cambridge University Press, Cambridge, 2001).

[65] E. Brézin and S. R. Wadia, *The Large N Expansion in Quantum Field Theory and Statistical Physics* (World Scientific, Singapore, 1993).

[66] M. L. Mehta, A method of integration over matrix variables, in *The Large N Expansion in Quantum Field Theory and Statistical Physics: From Spin Systems to 2-Dimensional Gravity* (World Scientific, Singapore, 1993), pp. 616–629.

[67] S. Chadha, G. Mahoux, and M. L. Mehta, A method of integration over matrix variables: 2, J. Phys. A: Math. Gen. **14**, 579 (1981).

[68] V. A. Kazakov, Solvable matrix models, arXiv:hep-th/0003064 (2000).

[69] V. Kazakov and A. Migdal, Induced gauge theory at large N, Nucl. Phys. B **397**, 214 (1993).

[70] V. Kazakov, D-dimensional induced gauge theory as a solvable matrix model, Nucl. Phys. B: Proc. Suppl. **30**, 149 (1993).

[71] B. Eynard, T. Kimura, and S. Ribault, Random matrices, arXiv:1510.04430 (2015).

[72] P. Zinn-Justin, The dilute Potts model on random surfaces, J. Stat. Phys. **98**, 245 (2000).

[73] M. Stephanov, J. Verbaarschot, and T. Wettig, Random matrices, arXiv:hep-ph/0509286 (2005).

[74] A. Zvonkin, Matrix integrals and map enumeration: An accessible introduction, Math. Comput. Model. **26**, 281 (1997).

[75] A. Matytsin and P. Zaugg, Kosterlitz-Thouless phase transitions on discretized random surfaces, Nucl. Phys. B **497**, 658 (1997).

[76] J. Barbier, Overlap matrix concentration in optimal Bayesian inference, Inf. Inference: J. IMA **10**, 597 (2021).

[77] A. Guionnet, First order asymptotics of matrix integrals: A rigorous approach towards the understanding of matrix models, Commun. Math. Phys. **244**, 527 (2004).

[78] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond*, Lecture Notes in Physics, Vol. 9 (World Scientific, Singapore, 1987).

[79] M. Mézard and A. Montanari, *Information, Physics and Computation* (Oxford University Press, Oxford, 2009).

[80] K. J. Johansson, Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices, Commun. Math. Phys. **215**, 683 (2001).

[81] N. Rosenzweig and C. E. Porter, "Repulsion of energy levels" in complex atomic spectra, Phys. Rev. **120**, 1698 (1960).

[82] E. Brézin and S. Hikami, Correlations of nearby levels induced by a random potential, Nucl. Phys. B **479**, 697 (1996).

[83] G. Biroli and M. Tarzia, Lévy-Rosenzweig-Porter random matrix ensemble, Phys. Rev. B **103**, 104205 (2021).

[84] V. Kravtsov, I. Khaymovich, E. Cuevas, and M. Amini, A random matrix model with localization and ergodic transitions, New J. Phys. **17**, 122002 (2015).

[85] D. Facoetti, P. Vivo, and G. Biroli, From non-ergodic eigenvectors to local resolvent statistics and back: A random matrix perspective, EPL (Europhys. Lett.) **115**, 47003 (2016).

[86] We require symmetry except in the trivial case $p_{S,N}(\boldsymbol{\lambda}^S) = \delta(\boldsymbol{\lambda}^S - \boldsymbol{\lambda}_0^S)$ for some fixed $\boldsymbol{\lambda}_0^S$.

[87] A. Guionnet and O. Zeitouni, Large deviations asymptotics for spherical integrals, J. Funct. Anal. **188**, 461 (2002).

[88] A. Matytsin, On the large-N limit of the Itzykson-Zuber integral, Nucl. Phys. B **411**, 805 (1994).

[89] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices*, Fields Institute Monographs, Vol. 35 (Springer, 2017).

[90] J. Pielaszkiewicz and M. Singull, Closed form of the asymptotic spectral distribution of random matrices using free independence, Report No. LiTH-MAT-R (Linkoping Univeristy, 2015).

[91] D. Guo, S. Shamai, and S. Verdú, Mutual information and minimum mean-square error in Gaussian channels, IEEE Trans. Info. Theory **51**, 1261 (2005).

[92] The factor 4 that differs from the 2 in the usual I-MMSE relation [91] comes from the fact that the Wigner matrix to denoise has only a fraction $N(N+1)/(2N^2) = 1/2 + O(1/N)$ of independent entries. The $O(1/N)$ correction comes from the diagonal terms in matrix $\boldsymbol{S}$ for which the signal-to-noise ratio is different than the one of the off-diagonal entries. The complex noise case of the I-MMSE relation is discussed in Sec. V D of [91].

[93] J. Fischmann, W. Bruzda, B. A. Khoruzhenko, H.-J. Sommers, and K. Życzkowski, Induced Ginibre ensemble of random matrices and quantum operations, J. Phys. A: Math. Theor. **45**, 075203 (2012).

[94] G. Akemann, J. R. Ipsen, and M. Kieburg, Products of rectangular random matrices: Singular values and progressive scattering, Phys. Rev. E **88**, 052118 (2013).

[95] J. R. Ipsen and M. Kieburg, Weak commutation relations and eigenvalue statistics for products of rectangular random matrices, Phys. Rev. E **89**, 032106 (2014).

[96] F. Hiai and D. Petz, A large deviation theorem for the empirical eigenvalue distribution of random unitary matrices, Ann. Inst. Henri Poincaré B **36**, 71 (2000).

[97] G. B. Arous and A. Guionnet, Large deviations for Wigner's law and Voiculescu's non-commutative entropy, Prob. Theory Related Fields **108**, 517 (1997).

[98] G. B. Arous and O. Zeitouni, Large deviations from the circular law, ESAIM: Prob. Stat. **2**, 123 (1998).

[99] Y. V. Fyodorov and I. Williams, Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity, J. Stat. Phys. **129**, 1081 (2007).

[100] J. Barbier and D. Panchenko, Strong replica symmetry in high-dimensional optimal Bayesian inference, arXiv:2005.03115 (2020).

[101] J. Leake, C. McSwiggen, and N. K. Vishnoi, Sampling matrices from Harish-Chandra–Itzykson–Zuber densities with applications to quantum inference and differential privacy, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (ACM digital library, 2021), pp. 1384–1397.

[102] J. Bun, J.-P. Bouchaud, S. N. Majumdar, and M. Potters, Instanton Approach to Large N Harish-Chandra-Itzykson-Zuber Integrals, Phys. Rev. Lett. **113**, 070201 (2014).

[103] J.-B. Zuber, The large-N limit of matrix integrals over the orthogonal group, J. Phys. A: Math. Theor. **41**, 382001 (2008).

[104] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, 2020).

[105] B. Collins, Moments and cumulants of polynomial random variables on unitarygroups, the Itzykson-Zuber integral, and free probability, Int. Math. Res. Notices **2003**, 953 (2003).

[106] P. Zinn-Justin and J.-B. Zuber, On some integrals over the $U(N)$ unitary group and their large N limit, J. Phys. A: Math. Gen. **36**, 3173 (2003).

[107] A. Edelman and N. R. Rao, Random matrix theory, Acta Numer. **14**, 233 (2005).

[108] B. Schlittgen and T. Wettig, Generalizations of some integrals over the unitary group, J. Phys. A: Math. Gen. **36**, 3195 (2003).

[109] A. Ghaderipoor and C. Tellambura, Generalization of some integrals over unitary matrices by character expansion of groups, J. Math. Phys. **49**, 073519 (2008).

[110] P. J. Forrester and J. Grela, Hydrodynamical spectral evolution for random matrices, J. Phys. A: Math. Theor. **49**, 085203 (2016).

[111] A. Guionnet and J. Huang, Large deviations asymptotics of rectangular spherical integral, arXiv:2106.07146 (2021).

[112] T. Tao and V. Vu, Random covariance matrices: Universality of local statistics of eigenvalues, Ann. Prob. **40**, 1285 (2012).

[113] P. J. Forrester and D.-Z. Liu, Singular values for products of complex Ginibre matrices with a source: Hard edge limit and phase transition, Commun. Math. Phys. **344**, 333 (2016).

[114] Z. Burda, A. Jarosz, G. Livan, M. A. Nowak, and A. Swiech, Eigenvalues and singular values of products of rectangular Gaussian random matrices, Phys. Rev. E **82**, 061114 (2010).

[115] J. R. Ipsen, Products of independent Gaussian random matrices, arXiv:1510.06128 (2015).

[116] J. R. Ipsen, Products of independent quaternion Ginibre matrices and their correlation functions, J. Phys. A: Math. Theor. **46**, 265201 (2013).

[117] G. Akemann and Z. Burda, Universal microscopic correlation functions for products of independent Ginibre matrices, J. Phys. A: Math. Theor. **45**, 465201 (2012).

[118] G. Akemann, M. Kieburg, and L. Wei, Singular value correlation functions for products of Wishart random matrices, J. Phys. A: Math. Theor. **46**, 275205 (2013).

[119] G. Akemann, Z. Burda, M. Kieburg, and T. Nagao, Universal microscopic correlation functions for products of truncated unitary matrices, J. Phys. A: Math. Theor. **47**, 255202 (2014).

[120] M. Kieburg, A. B. Kuijlaars, and D. Stivigny, Singular value statistics of matrix products with truncated unitary matrices, Int. Math. Res. Notices **2016**, 3392 (2016).

[121] G. Akemann and J. R. Ipsen, Recent exact and asymptotic results for products of independent random matrices, Acta Phys. Polon. B **46**, 1747 (2015).

[122] D.-Z. Liu, D. Wang, and L. Zhang, Bulk and soft-edge universality for singular values of products of Ginibre random matrices, Ann. Inst. H. Poincaré Probab. Statist. **52**, 1734 (2016).

[123] W. FitzGerald and N. Simm, Fluctuations and correlations for products of real asymmetric random matrices, arXiv:2109.00322 (2021).

[124] E. Gudowska-Nowak, R. A. Janik, J. Jurkiewicz, and M. A. Nowak, Infinite products of large random matrices and matrix-valued diffusion, Nucl. Phys. B **670**, 479 (2003).

[125] Z. Burda, R. A. Janik, and B. Waclaw, Spectrum of the product of independent random gaussian matrices, Phys. Rev. E **81**, 041132 (2010).

[126] F. Götze and A. Tikhomirov, On the asymptotic spectrum of products of independent random matrices, arXiv:1012.2710 (2010).

[127] S. O'Rourke and A. Soshnikov, Products of independent non-Hermitian random matrices, Elec. J. Prob. **16**, 2219 (2011).

[128] G. Akemann, J. Baik, and P. Di Francesco, *The Oxford Handbook of Random Matrix Theory* (Oxford University Press, Oxford, 2011).

[129] We note that this assumption is reminiscent of results in [132] (see also [133]) for the addition of two large random matrices with at least one being biunitary invariant.

[130] G. Menon, The complex Burgers equation, the HCIZ integral and the Calogero-Moser system (2017), http://www.dam.brown.edu/people/menon/talks/cmsa.pdf.

[131] W. Bryc, Computing moments of free additive convolution of measures, Appl. Math. Comput. **194**, 561 (2007).

[132] F. Benaych-Georges, Rectangular random matrices, related convolution, Probab. Theory Relat. Fields **144**, 471 (2009).

[133] S. T. Belinschi, F. Benaych-Georges, and A. Guionnet, Regularization by free additive convolution, square and rectangular cases, Complex Anal. Operator Theory **3**, 611 (2009).