OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **53** (2020) 164001 (14pp)
https://doi.org/10.1088/1751-8121/ab59ef

On the universality of noiseless linear estimation with respect to the measurement matrix

Alia Abbara[1], **Antoine Baker**[1], **Florent Krzakala**[1]
and Lenka Zdeborová[2]

[1] Laboratoire de Physique de l'Ecole normale supérieure, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité, Paris, France
[2] Institut de physique théorique, Université Paris Saclay, CNRS, CEA, 91191 Gif-sur-Yvette, France

E-mail: alia.abbara@ens.fr

Received 11 June 2019, revised 5 November 2019
Accepted for publication 21 November 2019
Published 31 March 2020

Abstract

In a noiseless linear estimation problem, the goal is to reconstruct a vector $\mathbf{x}^*$ from the knowledge of its linear projections $\mathbf{y} = \Phi\mathbf{x}^*$. There have been many theoretical works concentrating on the case where the matrix Φ is a random i.i.d. one, but a number of heuristic evidence suggests that many of these results are universal and extend well beyond this restricted case. Here we revisit this problem through the prism of development of message passing methods, and consider not only the universality of the ℓ_1-transition, as previously addressed, but also the one of the optimal Bayesian reconstruction. We observed that the universality extends to the Bayes-optimal minimum mean-squared (MMSE) error, and to a range of structured matrices.

Keywords: linear regression, random matrices, compressed sensing, replica method, approximate message passing

(Some figures may appear in colour only in the online journal)

1. Introduction

The problem of recovering a signal through the knowledge of its linear projections is ubiquitous in modern information theory, statistics and machine learning. In particular, many applications require to reconstruct an unknown $n-$ dimensional signal vector $\mathbf{x}^*$ from the linear projections

$$\mathbf{y} = \Phi\mathbf{x}^* \,, \tag{1}$$

where $\mathbf{y}$ is a m-dimensional vector, and Φ is an $m \times n$ random matrix. For instance, if $\mathbf{x}^*$ is sparse, this task of estimating the signal from its linear *random* projections is at the roots of compressed sensing [1]. A fundamental question in the field is how much the algorithmic and the information theoretic performance depends on the choice of the random matrix Φ.

In the present letter, we concentrate on the noiseless and asymptotic, large n, regime with a fixed value $\alpha = m/n$. We consider $\mathbf{x}^*$ to be k-sparse, i.e. to have only k non-zero values, and we shall work in the limit where $n \to \infty$, $k \to \infty$, and a finite value of $\rho = k/n$. In such setting, a classical result is the following: for random matrices Φ which entries are identically and independently generated from a Gaussian distribution, that we call Gaussian i.i.d. matrices, the (convex) reconstruction with ℓ_1-penalty displays a precisely determined phase transition. For a certain region in the (α, ρ)-phase diagram, it typically finds back the vector $\mathbf{x}^*$, being the sparsest solution, whereas outside that region, it typically fails. The boundary between these two regions is called the Donoho–Tanner line [2]. It has been shown empirically that the very same phase transition location seems to hold for a wider range of random matrix ensembles, see e.g. [3, 4], suggesting a large universality of the Donoho–Tanner phase transitions. Another line of work showed that the convex ℓ_1-reconstruction problem can be treated through conic geometry, and the success probability of signal recovery only depends on a geometric number characterizing a subcone (statistical dimension or Gaussian width) [5, 6].

Here we investigate the universality of the phase transition not only for the ℓ_1-transition, but also to the performance of the optimal Bayesian reconstruction. We analyze this question through the prism of information theory, message passing methods, and random matrix theory. We shall see that the universality indeed extends to a more generic set of properties than the ℓ_1-transition, such as the minimum mean-squared (MMSE) error or the easy-hard phase transition for optimal Bayesian learning, and empirically to structured matrices such as the one appearing in [7, 8].

We note that investigation of universality are very common to physics problems, and understanding how large is the class of model for which a given result applied is a very fundamental question. The message-passing-based algorithm that we investigate in this paper to demonstrate the universality also has their origin in physics works, such as [9].

2. A short review of results for independent and identically distributed (i.i.d.) random matrices

A first well-understood case of universality holds for random matrices Φ where all the elements are generated identically and independently from a well-behaved distribution -with zero mean and unit variance- which all exhibit the same transitions as Gaussian random matrices. This is known for multiple retrieval problems:

2.1. ℓ_1-recovery

Consider for instance the Donoho–Tanner line [2] that regulates the ℓ_1-recovery. Thanks to the approximate message passing solver (see below) that has been shown to be universal with respect to all matrices which are independent and identically distributed with finite moments [10, 11], we know that the Donoho–Tanner phase transition is the same for all those matrices.

2.2. Information theoretic optimal reconstruction

There has been a considerable amount of work in the information theory community on the computation of the mutual information and on the MMSE for problems such as (1) with Gaussian matrices. In particular, following the replica method from statistical physics (the Tanaka formula [12]), a heuristic formula has been postulated in different situations, see e.g. [13–16]. This heuristic replica result has been recently rigorously proven in a series of papers [17, 18]. In a more recent proof [19], it has been shown, again, that the formula is not specific to Gaussian i.i.d. matrices, but that any matrix with i.i.d elements of unit variance and zero mean leads to the same exact result for the mutual information and the MMSE.

2.3. Hard phase for Bayesian decoders

A third interesting point is to ask about tractable decoders that aim at performing the optimal Bayesian estimation, i.e. with a perfect prior knowledge on the distribution of $\mathbf{x}^*$. For simplicity, consider for instance the case where each element of $\mathbf{x}^*$ has been sampled from a Gauss–Bernoulli distribution:

$$x_i \sim (1-\rho)\delta(x) + \rho\mathcal{N}(0,1)\,.$$

$\rho \in [0,1]$ is the ratio of non-zero components of $\mathbf{x}^*$. In this case, the best known solver is again AMP, using a Bayesian decoder (instead of the soft thresholding function for ℓ_1-recovery) [14, 15, 20, 21]. Interestingly, it shares with the ℓ_1-recovery a similar phase transition: for a certain region in the (α,ρ) plane it typically finds back the vector $\mathbf{x}^*$, whereas outside that region it fails. We shall denote the limit between these regions the ‘Bayesian hard-phase’ transition. The ‘Bayesian hard-phase’ line, that has been precisely computed in [14, 15] is always better than the Donoho–Tanner line (as it should, since it exploits additional information). Once more, the universality of AMP shows that this phase transition is not restricted to Gaussian matrices, but extends as well to all (well normalized) i.i.d. matrices.

The fact that these three properties (the ℓ_1, the hard-phase line, as well as the MMSE) are universal for all i.i.d. matrices makes the case for Gaussian computations, as done in theoretical computation, stronger. We shall see that this universality extends well beyond these simple cases.

3. Random rotationally invariant matrices

Moving away from the well-known i.i.d. examples, we start by considering a much larger set of random matrices defined through their singular value decomposition (SVD): any real matrix Φ can be decomposed into $\Phi = U\Sigma V$, with U and V orthogonal matrices, and Σ's elements being Φ's singular values. We shall look at the left rotationally invariant random matrix ensemble: these are matrices Φ that can be written as

$$\Phi = U\Sigma V$$

with an arbitrary rotation matrix U and singular values Σ, but where the matrix V has been randomly (and independently of Σ and U) generated from the Haar measure (that is, uniformly from all possible rotations).

When the singular values are different from zero, it is straightforward to justify the universality property for matrices from this subclass. We start by the definition of the problem: we wish to find $\boldsymbol{x}$ such that

$$\boldsymbol{y} = \Phi\boldsymbol{x} = U\Sigma V\boldsymbol{x}. \tag{2}$$

If $m \leqslant n$, then Σ is written as $\Sigma = \left[\tilde{\Sigma} \mid 0\right]$ and we define

$$\Sigma^{\text{inv}} = \left[\begin{array}{c}\tilde{\Sigma}^{-1} \\ \hline 0\end{array}\right] \text{ such that } \Sigma^{\text{inv}}\Sigma = \left[\begin{array}{c|c} I_m & 0 \\ \hline 0 & 0\end{array}\right].$$

Multiplying (2) on both sides by U^T, and then by Σ^{inv}; one reaches

$$\tilde{\boldsymbol{y}} = \Sigma^{\text{inv}}U^T\boldsymbol{y} = \tilde{V}\boldsymbol{x} \tag{3}$$

where $\tilde{V}$ is an $m \times n$ matrix composed of the first m lines of V. If instead $m > n$, Σ is written as

$$\Sigma = \left[\begin{array}{c}\tilde{\Sigma} \\ \hline 0\end{array}\right]$$

and we define $\Sigma^{\text{inv}} = \left[\tilde{\Sigma}^{-1} \mid 0\right]$ such that $\Sigma^{\text{inv}}\Sigma = I_n$. Multiplying (2) by U^T then Σ^{inv}, we obtain a similar equation

$$\tilde{\boldsymbol{y}} = \Sigma^{\text{inv}}U^T\boldsymbol{y} = V\boldsymbol{x}. \tag{4}$$

In both cases, we thus see that the problem has been transformed—in a constructive way—into a standard linear system with the sensing matrix $\tilde{V}$ when $m \leqslant n$ being a (sub-sampled) random rotation one, or sensing matrix V when $m > n$. This shows that all rotationally invariant matrices, which satisfy U and Σ's independence on V, can be transformed the same way and are in the same universality class as far as noiseless linear recovery is concerned, i.e. they will display the same phase transitions.

Since Gaussian i.i.d. matrices belong among random rotationally invariant matrices (in this case Σ follows the Marcenko–Pastur law [22]) this means that all the information theoretic rigorous results (such as phase transitions and MMSE value) with zero noise for random Gaussian matrices applies verbatim to all rotationally invariant ensemble, as long as the SVD's matrices U and Σ are independent of V. This is a very strong universality, that applies to the phase transitions of the three cases (1, 2, 3) from section 2. Note that the universality of the Donoho–Tanner line with rotationally invariant matrices was already hinted by the replica method [23].

However, note that the above construction depends crucially on the fact that we consider here noiseless measurements. It would not work if an additional Gaussian noise were added in equation (1): in this case, the transformation would make the i.i.d. Gaussian noise a correlated one. Indeed, the replica formula for noisy measurements underlines that the MMSE depends on the precise set of matrices in noisy reconstruction [13, 24] (this formula is not yet fully rigorous, but see [25] for a proof in a restricted setting). Any differences, however, must go to zero in the noiseless limit.

4. Approximate message passing

Having discussed the universality with respect to random rotationally invariant matrices, we now wish to discuss its effect on specific solvers, concretely the message passing algorithms.

4.1. AMP

We first consider the original approximate message passing (AMP) [26] to compute the phase transition between the phase where the algorithm reconstructs $\mathbf{x}^*$ perfectly, and the one where reconstruction may be possible but is not achieved by the algorithm. AMP is an iterative algorihm that follows:

$$\begin{aligned}\hat{\mathbf{x}}^{t+1} &= \eta_t(\Phi^T\hat{\mathbf{x}}^t)\\ \mathbf{z}^t &= \mathbf{y} - \Phi\hat{\mathbf{x}}^t + \frac{1}{\alpha}\mathbf{z}^{t-1}\langle\eta'_{t-1}(\Phi^T\mathbf{z}^{t-1} + \hat{\mathbf{x}}^{t-1})\rangle.\end{aligned}$$

where t is the iteration index, $\mathbf{x}^t$ is the current estimate of $\mathbf{x}^*$, $\mathbf{z}^t$ the current residual, $\langle\cdot\rangle$ is an averaged sum of components, and η_t is a prior-dependent threshold function applied component-wise (the soft thresholding for ℓ_1, or the Bayesian decoder [14, 15]).

One of the most interesting features of AMP is that, if Φ is a Gaussian i.i.d. matrix, its mean squared error (MSE) σ_t can be tracked accurately by the state evolution formalism [10, 11, 26]. State evolution is a relatively simple recursive equation:

$$\sigma_{t+1}^2 = \Psi(\sigma_t^2)\,,\ \Psi(\sigma^2) = \mathbb{E}\left[\left(\eta_t(X + \frac{\sigma}{\sqrt{\alpha}}Z) - X\right)^2\right]\,, \tag{5}$$

where the expectation is with respect to independent random variables $Z \sim \mathcal{N}(0, 1)$ and X, whose distribution coincides with the empirical distribution of the entries of x^*. Analyzing the evolution of this equation for the ℓ_1-decoder yields the Donoho–Tanner line [26], while using the Bayesian decoder it yields the hard-phase line for Bayesian decoding [14].

It would be interesting to use AMP for rotationally invariant matrices. In order to do this, we follow the construction of section 3: starting from equation (3) we then multiply by Σ_0, an $m \times m$ diagonal matrix with singular values sampled from Marcenko–Pastur law (singular values of a Gaussian i.i.d. matrix[3]), and U_0 an $m \times m$ Haar-generated orthogonal matrix, thus ensuring that Σ_0 and U_0 are generated independently of V:

$$U_0\Sigma_0\tilde{\Sigma}^{-1}U^T\boldsymbol{y} = U_0\Sigma_0\tilde{V}\boldsymbol{x} \tag{6}$$

$$\boldsymbol{y}' = \Phi'\boldsymbol{x}. \tag{7}$$

After this transformation, $\Phi' = U_0\Sigma_0\tilde{V}$ is a random matrix that belongs to an ensemble very close to the Gaussian i.i.d. matrices ensemble. In fact, a recent work showed that AMP applied to a Gaussian matrix follows the same state evolution as matrices such as Φ' where $U_0, \tilde{V}$ are uniform orthogonal matrices and Σ_0 diagonal's elements are singular values sampled from the Marcenko–Pastur law [27]. Combining this result with the matrix transformation, we have thus constructively mapped the noiseless reconstruction problem back to the well-understood noiseless compressed sensing case for a Gaussian i.i.d. matrix, where we can safely apply the

---

[3] The singular values of a Gaussian matrix are correlated, so in fact we may want to generate Σ_0 by first generating a random Gaussian matrix, and then calculating its singular values.

algorithm, and its state evolution. In the section 5.2, we apply this matrix transformation for numerical experiments using AMP.

4.2. Vector-AMP

While the transformation trick allows to make AMP work with random rotationally invariant matrices, another alternative is to work directly with a dedicated solver. To this means, different but related approaches were proposed [24, 28], in particular, using the general expectation-propagation (EP) [29, 30] scheme. Ma and Ping proposed a variation of EP called OAMP [31] specially adapted to rotation matrices. Rangan, Schniter and Fletcher introduced a similar approach called VAMP [32] and proved that it follows state evolution equations corresponding to the fixed point of the replica potential [13, 24, 25]. The multi-layer AMP algorithm of [33] also display the same fixed point.

We shall concentrate here on the VAMP (Vector-AMP) approach, and for a moment, put back a small additional random Gaussian i.i.d. noise of variance Δ in the measurement in equation (1) as it is needed for stating the algorithm. VAMP then consists in the following fixed-point iteration:

$$\begin{aligned} \boldsymbol{u}_\ell^{t+1} &= \frac{\hat{\boldsymbol{x}}_{\boldsymbol{l}}{}^t}{\langle \mathrm{Var}_\ell^t(\boldsymbol{x})\rangle} - \boldsymbol{u}_r^t, \qquad & \rho_\ell^{t+1} &= \frac{1}{\langle \mathrm{Var}_\ell^t(\boldsymbol{x})\rangle} - \rho_r^t, \\ \boldsymbol{u}_r^{t+1} &= \frac{\hat{\boldsymbol{x}}_{\boldsymbol{r}}{}^t}{\langle \mathrm{Var}_r(\boldsymbol{x})\rangle} - \boldsymbol{u}_\ell^t, \qquad & \rho_r^{t+1} &= \frac{1}{\langle \mathrm{Var}_r^t(\boldsymbol{x})\rangle} - \rho_\ell^t, \end{aligned} \tag{8}$$

where $\mathbb{E}_{\ell,r}^t$ and $\mathrm{Var}_{r,\ell}^t(\boldsymbol{x})$ are the expectation and variance of the tilted distributions

$$\tilde{Q}_{\ell,r}^t(\boldsymbol{x}) \propto P_{\ell,r}(\boldsymbol{x}) Q_{\ell,r}^t(\boldsymbol{x}), \tag{9}$$

where

$$Q_{l,r}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\rho_{l,r}\boldsymbol{x}^T\boldsymbol{x} + \boldsymbol{u}_{l,t}^T\boldsymbol{x}} \tag{10}$$

$$P_l(\boldsymbol{x}) \propto \mathrm{e}^{-||\mathbf{y}-\Phi\mathbf{x}||_2^2/2\Delta}, \tag{11}$$

and $P_r(\boldsymbol{x})$ is the prior used in the algorithm (i.e. the Laplace prior for the ℓ_1-model, or the actual distribution of the signal for Bayesian reconstruction). In particular

$$\begin{aligned} \hat{\boldsymbol{x}}_{\boldsymbol{l}}{}^t &= (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \Delta\rho_r^t \boldsymbol{I}_p)^{-1}(\boldsymbol{\Phi}^T\boldsymbol{y} + \Delta\boldsymbol{u}_r^t), \\ \langle \mathrm{Var}_\ell^t(\boldsymbol{x})\rangle &= \frac{\Delta}{N}\mathrm{Tr}(\boldsymbol{\phi}^T\boldsymbol{\Phi} + \Delta\rho_r^t \boldsymbol{I}_p)^{-1}, \end{aligned} \tag{12}$$

where, as for AMP, we define the denoiser that yields the estimates of x by $z(u,\rho) = \int dx P_r(x)\mathrm{e}^{-\frac{1}{2}\rho x^2 + ux}$,

$$\begin{aligned} (\hat{x}_r)_j &= \frac{\partial}{\partial u}\log z(u,\rho)\Big|_{u_{\ell k}^t,\rho_\ell^t}, \\ \langle \mathrm{Var}_r^t(\boldsymbol{x})\rangle &= \frac{1}{n}\sum_{j=1}^{p}\frac{\partial^2}{\partial u^2}\log z(u,\rho)\Big|_{u_{\ell k}^t,\rho_\ell^t}. \end{aligned} \tag{13}$$

Again, the performance of the recursion can be analyzed rigorously through the state evolution [32]. For simplicity, let us concentrate on the Bayes optimal case in which case the state evolution can be closed on the variables (see [32]):

$$\sigma^t = \langle \mathrm{Var}_r^t(\boldsymbol{x}) \rangle \text{ and } \epsilon^t = \langle \mathrm{Var}_l^t(\boldsymbol{x}) \rangle \,, \tag{14}$$

by writing

$$\begin{aligned} \sigma^t(\rho_l^t) &= \Psi((\rho_l^t)^{-1}) \\ \epsilon^t(\rho_r^t) &= \frac{\Delta}{N} \mathrm{Tr}\left[\left(\Sigma^2 + \Delta\rho_r^t I_N\right)^{-1}\right] = \Delta S_{\Sigma^2}(-\Delta\rho_r^t) \end{aligned} \tag{15}$$

where I_N is the identity matrix of size N, and we recognize the Stieltjes transform $S_{\Sigma^2}(t) = \frac{1}{N}\mathrm{Tr}\left[\left(\Sigma^2 - tI_N\right)^{-1}\right]$.

Though this transform, we see that the performance depends crucially on the distribution of eigenvalues. Let us now go back on the noiseless limit when $\Delta \to 0$ and analyze how the universality shows up. Consider again the Stieltjes transform: out of the n singular values of the $n \times n$ matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$, we shall have $(1-\alpha)n$ of them to be zero (assuming $\alpha < 1$) while the rest are positive (since $m < n$). In this case, the limit $r \to 0$ of the Stieltjes transform will behave as $S_X(r) \approx -(1-\alpha)/r$ so that

$$\lim_{\Delta \to 0} \epsilon(\rho_r^t) = \frac{1-\alpha}{\rho_r^t} \,.$$

Again, we see that all the complicated dependence on the spectrum of the matrix Φ has been eliminated. This is a direct, alternative, proof that VAMP will also yield universal results in the zero noise limit for the Bayesian reconstruction. Given that VAMP has the same fixed point as the replica mutual information [13, 25], this argument applies to the replica prediction for the MMSE as well.

5. Structured matrices

We now move to very structured matrices, in order to test the universality as well as the quality and the prediction of the state evolution out of its comfort zone. In order to do so, we have considered different matrix ensembles:

5.1. Tested ensembled of matrices

5.1.1. Discrete cosine transform matrices. The first ensemble we consider consists in Fourier-like matrices. An $n \times n$ discrete cosine transform (DCT) matrix Y is defined by:

$$Y_{jk} = \sqrt{\frac{2}{n}} \epsilon_k \cos\left(\frac{\pi(2j+1)k}{2n}\right) \,, \tag{16}$$

where $j, k \in [\![0, n-1]\!]$, $\epsilon_0 = 1/\sqrt{2}$, $\epsilon_i = 1$ for $i = 1, ..., n-1$. We used a sub-sampled version of these matrices in which we picked some rows randomly.

5.1.2. Hadamard matrices. A natural variant of DCT is given by the Hadamard matrices. H is an $n \times n$ Hadamard matrix if its entries are ± 1 and its rows are pairwise orthogonal, i.e. $HH^T = nI_n$. For every integer k, there exists a Hadamard matrix H_k of size 2^k. These can be created with Sylvester's construction: let H be a Hadamard matrix of order n. Then the partitioned matrix

$$\begin{bmatrix} H & H \\ H & -H \end{bmatrix}$$

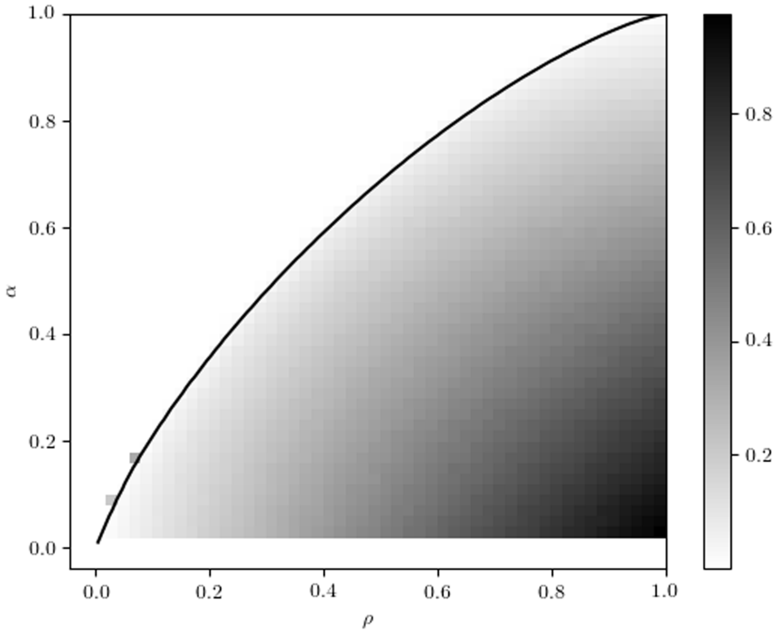

Figure 1. Phase diagram for a DCT matrix (width $n = 1000$) in the Bayes-optimal case. The averaged MSE on 50 executions of VAMP is represented by a color-code, displaying a phase transition that matches the theoretical Bayes line for Gaussian i.i.d. matrices (black line). Some finite-size effects can be seen.

is a Hadamard matrix of order $2n$.

5.1.3. Random features maps. Finally, we wanted to consider here random features maps (RFM) as encountered in nonlinear regression problems. In such settings, a random features matrix $\Phi = f(WX)$ is obtained from the raw data matrix X by means of a random projection matrix W and a pointwise nonlinear activation f. Kernel regression models, nonlinear in the original data X, can then be approximately but efficiently solved by the linear estimation problem (1), with an appropriate choice for f and the W-distribution [34]. Such matrices, that can be seen as the output of a neuron with random weights, have been investigated in particular in the context of neural networks [7, 8]. Indeed, in neural networks configurations with random weights play an important role as they define the initial loss landscape. They are also fundamental in the random kitchen sinks algorithm in machine learning [34] and it is thus of interest to test our understanding of linear reconstructions with AMP and VAMP in this case.

In what follows we will test random features matrices where both W and X are random Gaussian i.i.d. matrices.

5.2. Numerical results

We provide the codes used to generate the data on github in the repository http://sphinxteam/Universality-CS-2019. To generate figures 1 and 2, we ran VAMP 50 times on 50×50 points spanning the (α,ρ)-space, and computed the average mean-squared error (MSE) between the signal $\mathbf{x}^*$ and the reconstructed configuration $\mathbf{x}$. The MSE is represented with a color bar (white means perfect reconstruction). For a DCT and a Hadamard matrix, we observe a

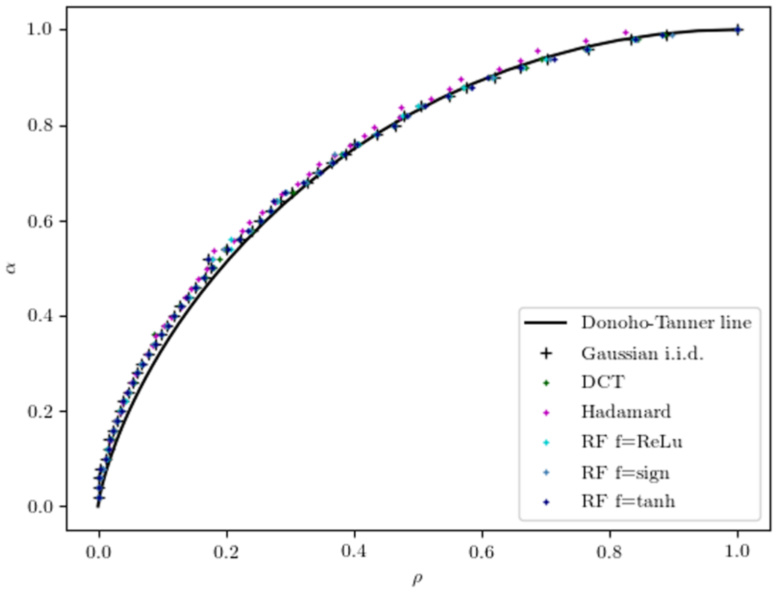

Figure 2. Phase diagram in the ℓ_1-reconstruction case obtained by averaging on 20 to 50 executions on VAMP. The dots indicate the phase transitions for Gaussian i.i.d., DCT (width $n = 2000$), Hadamard matrices ($n = 4096$); and random feature matrices $\Phi = f(WX)$ with $f = \mathrm{ReLu}$ ($\mathrm{ReLu}(x) = 0$ if $x \leqslant 0$, x if $x > 0$), $f = \mathrm{sign}$, $f = \tanh$ (W and X are Gaussian i.i.d. of size $\alpha n \times n$ and $n \times n$ with $n = 2000$). They match the theoretical Donoho–Tanner transition for Gaussian i.i.d. matrices (black line).

phase transition in the Bayes-optimal case that matches the theoretical transition for Gaussian i.i.d. matrices. We also ran VAMP for the ℓ_1-reconstruction problem. Averaging on 20 executions (or 50 for small α where finite-size effects are more important), we recover again a phase transition matching the theoretical Donoho–Tanner line for Gaussian i.i.d. matrices [3]. Besides, we compared the MSE obtained by VAMP at each point of the phase diagram for different matrices. In figures 3 and 4, we plot the MSE averaged on 20 executions for ρ fixed and α ranging between 0 and 1. We get the same error in reconstruction for all matrices, following the MSE for Gaussian i.i.d. matrix for $\rho = 0.25$, 0.5 and 0.75. We also checked that AMP, provided one uses the trick equation (7), reproduce these results as well: indeed the two algorithms returned extremely similar results.

5.3. Discussion

Figures of the previous section perfectly illustrate our main point: the universality in noiseless compressed sensing is not limited to the ℓ_1-type reconstruction as in [3, 4], but extends to other quantities and estimators, such as the hard-phase line in Bayesian reconstruction, and the MMSE. Besides, it is not limited to random orthogonal matrices, but empirically extends to Fourier-type matrices and to the random features maps currently studied in machine learning. It is an open question to extend the proof of state evolution to these challenging matrices. However, all matrices do not share the same properties of phase transitions and MMSE. Let us have a look at two examples of structured matrices that do not seem to follow these universal phase transitions.

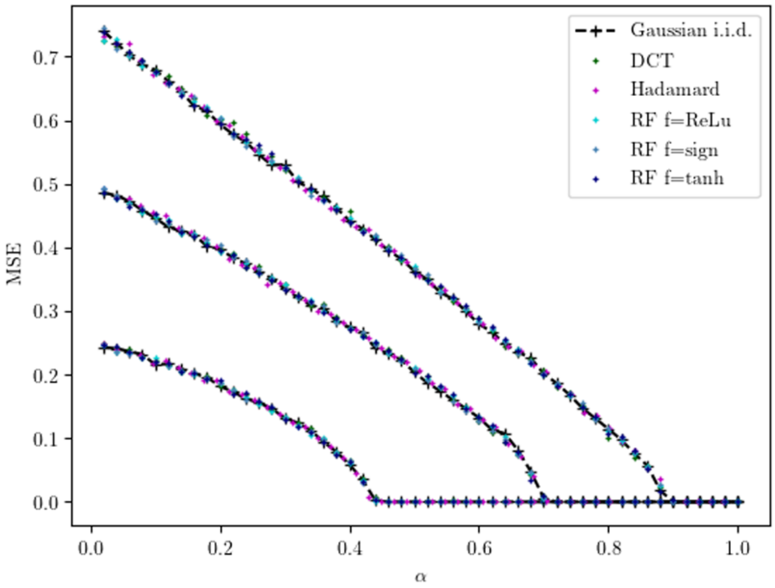

Figure 3. Mean-squared error for $\rho = 0.25$, 0.5 and 0.75 (bottom to up curves) in the Bayes-optimal case averaged on 20 executions of VAMP for Gaussian i.i.d, DCT, Hadamard, random features matrices $\Phi = f(WX)$ with $f = \text{ReLu}$, $f = \text{sign}$, $f = \tanh$ (W and X are Gaussian i.i.d of size $\alpha n \times n$ and $n \times n$) . The width is $n = 2000$ for all matrices.

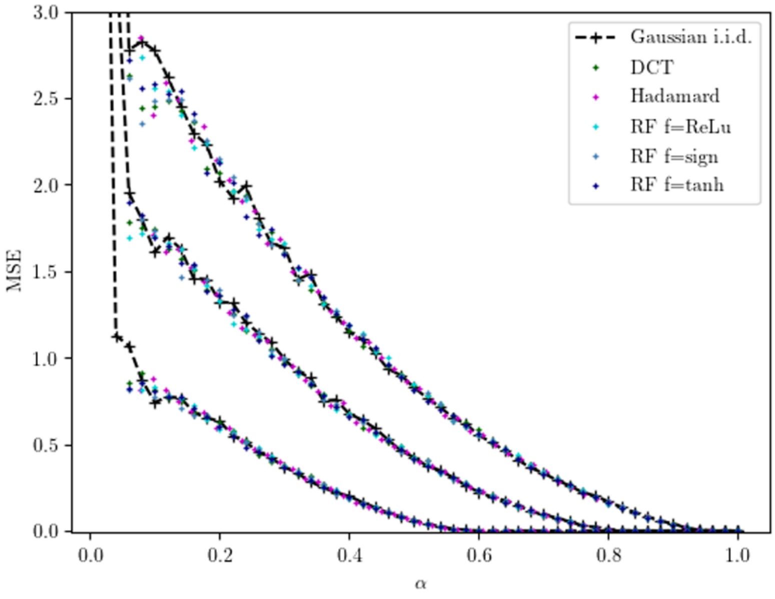

Figure 4. Mean-squared error for $\rho = 0.25$, 0.5 and 0.75 (bottom to up curves) in the ℓ_1-reconstruction case averaged on 20 executions of VAMP for Gaussian i.i.d, DCT, Hadamard, random features matrices $\Phi = f(WX)$ with $f = \text{ReLu}$, $f = \text{sign}$, $f = \tanh$ (W and X are Gaussian i.i.d of size $\alpha n \times n$ and $n \times n$). The width is $n = 2000$ for all matrices.

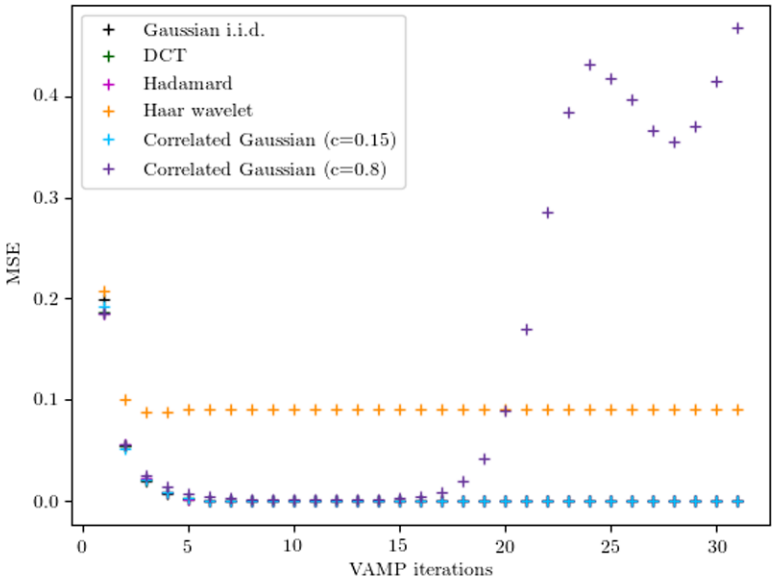

Figure 5. Mean-squared error at each iteration of VAMP in a Bayes-optimal setting, for $\rho = 0.3$, $\alpha = 0.7$ (in the easy phase of compressed sensing, i.e. below the 'Bayesian hard-phase' line) . VAMP is applied to a Gaussian i.i.d, a DCT, correlated Gaussian (width $n = 2000$); Hadamard and Haar wavelet matrices (width $n = 2048$). The MSE for the Haar wavelet matrix converges to a finite value but does not go to zero as for the other matrices. The MSE for a Gaussian correlated matrix converges for small correlation $c = 0.15$ and diverges for larger correlation $c = 0.8$.

- Haar wavelet matrices

Haar wavelet matrices can be defined recursively by:

$$W_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and } W_{2k} = \begin{bmatrix} H_k \otimes [1, -1] \\ I_k \otimes [1, 1] \end{bmatrix}$$

where I_k is the identity matrix of size k and $\otimes$ is the Kronecker product. In the easy phase of compressed sensing, both in the Bayes-optimal setting and the ℓ_1-recovery case, where VAMP applied to Gaussian i.i.d. matrices (as well as Hadamard, DCT, random features matrices) perfectly reconstructs the signal; it fails to do so when applied to a Haar wavelet matrix. VAMP will then converge to a fixed point with a non-zero MMSE, as seen in figures 5 and 6. In fact, VAMP seems to always fail in reconstructing the signal for a Haar wavelet matrix: the MMSE converges to a finite quantity, but never to zero. Hence we do not observe the same phase transitions for VAMP applied to a Haar wavelet matrix.

- Gaussian correlated matrices

Let $T(c)$ be the Toeplitz matrix defined as $T(c)_{ab} = c^{|a-b|}$. As in [35], we consider structured matrices which satisfy the following property: if M is a $m \times n$ matrix, its elements have covariance

$$\mathbb{E}[M_{ia}M_{jb}] = \frac{1}{m} C_{ij} D_{ab} \tag{17}$$

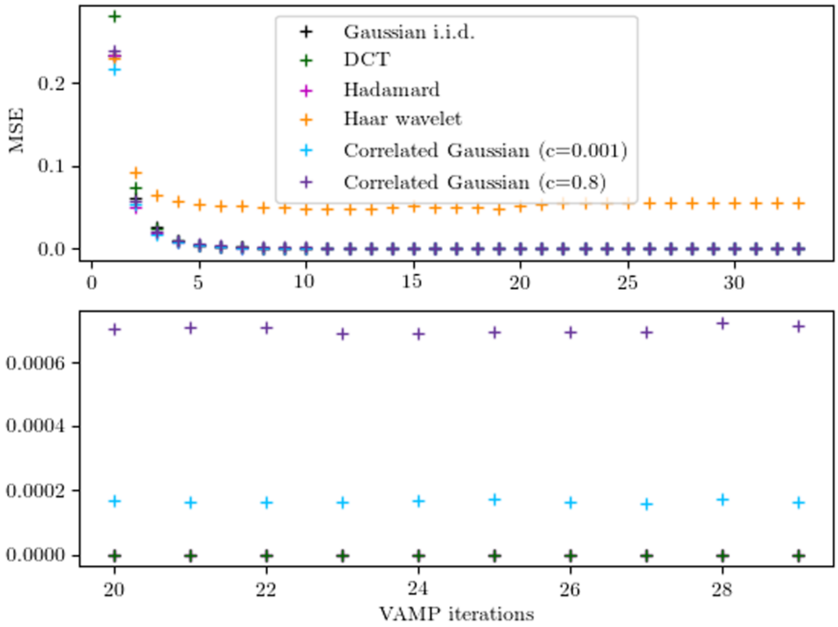

Figure 6. Mean-squared error at each iteration of VAMP in the case of ℓ_1-recovery, for $\rho = 0.3$, $\alpha = 0.8$ (in the easy phase of compressed sensing, below the Donoho–Tanner line). VAMP is applied to a Gaussian i.i.d, a DCT, correlated Gaussian (width $n = 2000$); Hadamard and Haar wavelet matrices (width $n = 2048$). The MSE for the Haar wavelet matrix converges to a finite value but does not go to zero. The MSE for Gaussian correlated matrices does not effectively converge, but stays very close to a small non-zero value, as seen in the zoomed-in second subplot.

where all $D_{aa} = 1$. Such a matrix can be obtained, for instance, by multiplying a $m \times n$ Gaussian i.i.d. matrix G by a $n \times n$ Toeplitz matrix $T(\sqrt{c})$. In our simulations, we thus used matrices

$$M(c) = \frac{1}{\sqrt{m}} G T(\sqrt{c}) \tag{18}$$

for different values of c. Running VAMP in the Bayes-optimal case with parameters (α, ρ) in the easy phase of compressed sensing, we find that it converges and perfectly reconstructs the signal for c small enough ($c = 0.15$), but fails to converge and has a diverging MSE when c is larger ($c = 0.8$), as seen in figure 5. In the ℓ_1-recovery setting, still in the easy phase of compressed sensing, VAMP fails to converge to a fixed reconstructed vector $\hat{\boldsymbol{x}}$ both for c very small ($c = 0.001$) or large ($c = 0.8$). However, the MSE stays very close to a small non-zero value, which can be seen in figure 6. After a large number of iterations, VAMP keeps returning a vector very close to the original signal, but does not manage to reconstruct it. The final MSE's approximate value also depends on c: the larger the correlations are, the larger the MSE is. In [35], the authors study such correlated matrices in the very sparse regime when α is close to zero, and show that the theoretical phase transition for ℓ_1-recovery depends on c.

Investigating the behaviour of these matrices in the non-sparse regime is an interesting direction of research. For now, it is clear that VAMP used on Haar wavelet matrices and Gaussian correlated matrices does not display the same phase transitions as Gaussian i.i.d., DCT, Hadamard and Random features matrices. It would be interesting to find a good criterion to identify which matrices satisfy this universality and which do not; this is something that we are yet unable to predict in advance.

Acknowledgment

We thank Andre Manoel and Galen Reeves for useful discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe; and from the French National Research Agency (ANR) Grant PAIL.

ORCID iDs

Alia Abbara https://orcid.org/0000-0002-5353-8993

References

[1] Candes E J and Tao T 2006 Near-optimal signal recovery from random projections: universal encoding strategies *IEEE Trans. Inf. Theory* **52** 5406–25

[2] Donoho D L and Tanner J 2005 Sparse nonnegative solution of underdetermined linear equations by linear programming *Proc. Natl Acad. Sci.* **102** 9446–51

[3] Donoho D and Tanner J 2009 Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing *Phil. Trans.* A **367** 4273–93

[4] Monajemi H, Jafarpour S, Gavish M and Donoho D L 2013 Deterministic matrices matching the compressed sensing phase transitions of gaussian random matrices *Proc. Natl Acad. Sci.* **110** 1181–6

[5] Chandrasekaran V, Recht B, Parrilo P A and Willsky A S 2012 The convex geometry of linear inverse problems *Found. Comput. Mathe.* **12** 805–49

[6] Amelunxen D, Lotz M, McCoy M B and Tropp J A 2014 Living on the edge: phase transitions in convex programs with random data *Inf. Inference* A **3** 224–94

[7] Pennington J and Worah P 2017 Nonlinear random matrix theory for deep learning *Advances in Neural Information Processing Systems* pp 2637–46

[8] Liao Z and Couillet R 2018 On the spectrum of random features maps of high dimensional data *Int. Conf. on Machine Learning* pp 3069–77 (http://proceedings.mlr.press/v80/liao18a.html)

[9] Thouless D J, Anderson P W and Palmer R G 1977 Solution of'solvable model of a spin glass *Phil. Mag.* **35** 593–601

[10] Bayati M and Montanari A 2011 The dynamics of message passing on dense graphs, with applications to compressed sensing *IEEE Trans. Inf. Theory* **57** 764–85

[11] Bayati M *et al* 2015 Universality in polytope phase transitions and message passing algorithms *Ann. Appl. Probab.* **25** 753–822

[12] Tanaka T 2001 Statistical mechanics of CDMA multiuser demodulation *Europhys. Lett.* **54** 540

[13] Tulino A M, Caire G, Verdu S and Shamai S 2013 Support recovery with sparsely sampled free random matrices *IEEE Trans. Inf. Theory* **59** 4243–71

[14] Krzakala F, Mézard M, Sausset F, Sun Y and Zdeborová L 2012 Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices *J. Stat. Mech.* P08009

[15] Krzakala F, Mézard M, Sausset F, Sun Y and Zdeborová L 2012 Statistical-physics-based reconstruction in compressed sensing *Phys. Rev.* X **2** 021005

[16] Zhu J and Baron D 2013 Performance regions in compressed sensing from noisy measurements *47th Annual Conf. on Inf. Sciences and Systems* (IEEE) pp 1–6

[17] Barbier J, Dia M, Macris N and Krzakala F 2016 The mutual information in random linear estimation *54th Annual Allerton Conf. on Communication, Control, and Computing (Allerton)* pp 625–32

[18] Reeves G and Pfister H D 2016 The replica-symmetric prediction for compressed sensing with gaussian matrices is exact *IEEE International Symp. on Information Theory* pp 665–9

[19] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 Optimal errors and phase transitions in high-dimensional generalized linear models *Proc. Natl Acad. Sci.* **116** 5451–60

[20] Vila J and Schniter P 2011 Expectation-maximization bernoulli-gaussian approximate message passing *Conf. Record of the 45th Asilomar Conf. on Signals, Systems and Computers* (IEEE) pp 799–803

[21] Montanari A 2012 Graphical models concepts in compressed sensing *Compressed Sensing: Theory and Applications* (Cambridge: Cambridge University Press) pp 394–438

[22] Tulino A M *et al* 2004 Random matrix theory and wireless communications *Found. Trends® Commun. Inf. Theory* **1** 1–182

[23] Kabashima Y, Wadayama T and Tanaka T 2009 A typical reconstruction limit for compressed sensing based on lp-norm minimization *J. Stat. Mech.* L09003

[24] Takeda K, Uda S and Kabashima Y 2006 Analysis of CDMA systems that are characterized by eigenvalue spectrum *Europhys. Lett.* **76** 1193

[25] Barbier J, Macris N, Maillard A and Krzakala F 2018 The mutual information in random linear estimation beyond i.i.d. matrices *IEEE Int. Symp. on Information Theory*

[26] Donoho D L, Maleki A and Montanari A 2009 Message-passing algorithms for compressed sensing *Proc. Natl Acad. Sci.* **106** 18914–9

[27] Takeuchi K 2019 A unified framework of state evolution for message-passing algorithms *IEEE Int. Symp. Inf. Theor.* (https://doi.org/10.1109/ISIT.2019.8849321)

[28] Cakmak B, Winther O and Fleury B H 2014 S-amp: approximate message passing for general matrix ensembles *IEEE Information Theory Workshop* (IEEE) pp 192–6

[29] Minka T P 2001 Expectation propagation for approximate bayesian inference *Proc. of the Seventeenth Conf. on Uncertainty in Artificial Intelligence* UAI'01 pp 362–9

[30] Opper M and Winther O 2005 Expectation consistent approximate inference *J. Mach. Learn. Res.* **6** 2177

[31] Ma J and Ping L 2017 Orthogonal amp *IEEE Access* **5** 2020–33

[32] Rangan S, Schniter P and Fletcher A K 2017 Vector approximate message passing *IEEE Int. Symp. on Information Theory* (IEEE) pp 1588–92

[33] Manoel A, Krzakala F, Mézard M and Zdeborová L 2017 Multi-layer generalized linear estimation *IEEE Int. Symp. on Information Theory* pp 2098–102

[34] Rahimi A and Recht B 2008 Random features for large-scale kernel machines *Advances in Neural Information Processing Systems* pp 1177–84

[35] Ramezanali M, Mitra P P and Sengupta A M 2016 Mean field analysis of sparse reconstruction with correlated variables *2016 24th European Signal Processing Conf.* (EUSIPCO) pp 1267–71