Phase retrieval in high dimensions: phase transitions and optimal spectral methods

arXiv:2006.05228, *NeurIPS '20*

arXiv:2012.04524, *MSML '21*

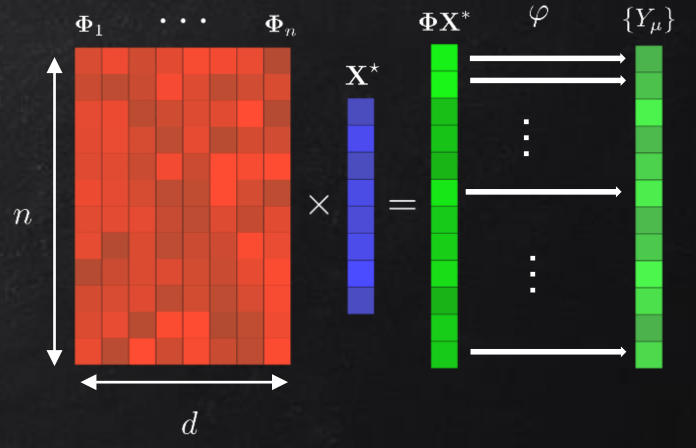
Antoine Maillard

Joint work with Bruno Loureiro, Florent Krzakala, Yue M. Lu & Lenka Zdeborová

Rigorous Evidence for Information-Computation Trade-offs – September 15th 2021

Phase retrieval

Goal : Recover a d-dimensional signal $\mathbf{X}^\star$ from n data points $\{\boldsymbol{\Phi}_\mu, Y_\mu\}_{\mu=1}^n$ generated as :



Generalized Linear Model (GLM)

Observations $Y_\mu \in \mathbb{R}$

$$\boxed{Y_\mu \sim P_{\text{out}}\left(\cdot \middle| \frac{1}{\sqrt{d}} \sum_{i=1}^d \Phi_{\mu i} X_i^\star\right) \quad \mu \in \{1, \cdots, n\}}$$

(Probabilistic) channel with possible noise.

Sensing matrix (real/complex)

Signal (real/complex), d-dimensional

Real / Complex

$\beta = 1$ $\quad$ $\beta = 2$

Phase retrieval: $P_{\text{out}}(y|z) = P_{\text{out}}(y||z|)$, e.g. noiseless $Y_\mu = \frac{1}{d}|(\boldsymbol{\Phi}\mathbf{X}^\star)_\mu|^2$; Poisson-noise $Y_\mu \sim \text{Pois}(\Lambda|(\boldsymbol{\Phi}\mathbf{X}^\star)_\mu|^2/d)$.

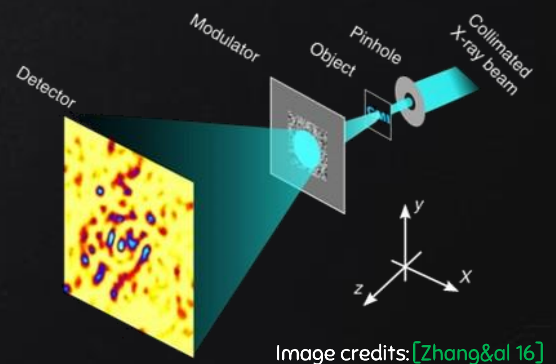


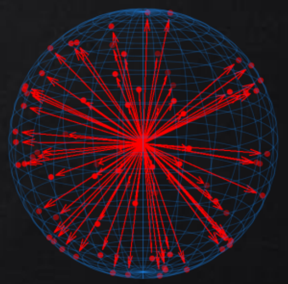
Image credits: [Zhang&al 16]

In the limit $d, n \to \infty$ with $\alpha = n/d = \Theta(1)$, what is the smallest α needed to recover $\mathbf{X}^\star$ ...

- Better than a random guess ?
- Perfectly ? (up to the possible rank deficiency of $\boldsymbol{\Phi}$)
- With which (polynomial-time) algorithm ? Cheap (e.g. spectral) methods ?

Goal : Fundamental limits of high-dimensional phase retrieval with random sensing matrix and signal in the typical case.

⚠ Different from “worst-case” injectivity studies [Bandeira&al '14]



Our model: The matrix $\mathbf{\Phi}$ is right-orthogonally (unitarily) invariant, i.e. delocalized right-eigenvectors: $\forall \mathbf{U},\ \mathbf{\Phi} \stackrel{d}{=} \mathbf{\Phi}\mathbf{U}$

The bulk of eigenvalues of $\mathbf{\Phi}^\dagger\mathbf{\Phi}/d$ converges to a distribution $\nu(x)$, as $n, d \to \infty$ with $n/d \to \alpha > 0$.

Examples: Gaussian matrices, product of Gaussians, random column-orthogonal/unitary, any $\mathbf{\Phi} \equiv \mathbf{USV}^\dagger$ with $S_i^2 \stackrel{\text{i.i.d.}}{\sim} \nu$.

The *Minimal Mean Squared Error (MMSE)* estimator is the first moment of the posterior distribution:

$$P(\mathrm{d}\mathbf{x}|\mathbf{Y},\mathbf{\Phi}) \equiv \frac{1}{\mathcal{Z}_d(\mathbf{Y},\mathbf{\Phi})}\prod_{i=1}^{d} P_0(\mathrm{d}x_i)\prod_{\mu=1}^{n} P_{\text{out}}\Big(Y_\mu\Big|\frac{1}{\sqrt{d}}\sum_{i=1}^{d}\Phi_{\mu i}x_i\Big)$$

Conjecture ("Replica formula"): $$\lim_{d\to\infty}\frac{1}{d}\mathbb{E}\ln \mathcal{Z}_d(\mathbf{Y},\mathbf{\Phi}) = \sup_{q_x,q_z}\overbrace{\Big[\underset{P_0}{I_0^{(\beta)}(q_x)} + \underset{P_{\text{out}}}{I_{\text{out}}^{(\beta)}(q_z)} + \underset{\nu}{\beta I_{\text{int}}(q_x,q_z)}\Big]}^{\text{“Replica-symmetric” potential } f(q_x,q_z)}$$

The *information-theoretic MMSE is*: $\lim_{d\to\infty}\mathbb{E}\|\mathbf{X}^\star - \hat{\mathbf{X}}_{\text{opt}}\|^2/d = \rho - q_x.$

- The functions involved in the optimization problem are fully explicit.
- The log-partition (or *free entropy*) is related to the *mutual information* $\mathrm{I}(\mathbf{X}^\star;\mathbf{Y}|\mathbf{\Phi}) = \mathbb{E}\ln\mathcal{Z}_d - n\mathbb{E}\ln P_{\text{out}}\big(Y_1\big|\frac{(\mathbf{\Phi}\mathbf{X}^\star)_1}{\sqrt{d}}\big)$
- Conjecture obtained with the heuristic replica method of statistical physics. [Mézard&al 1987, Takahashi&al '20] $\mathbb{E}\ln\mathcal{Z} = \lim_{r\to 0^+}\frac{\mathbb{E}\mathcal{Z}^r - 1}{r}$

Rigorous fundamental limits

Conjecture ("Replica formula"): $\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\ln \mathcal{Z}_d(\mathbf{Y},\boldsymbol{\Phi}) = \sup_{q_x,q_z} [I_0^{(\beta)}(q_x) + I_{\text{out}}^{(\beta)}(q_z) + \beta I_{\text{int}}(q_x,q_z)]$

The *information-theoretic MMSE is*: $\lim_{d\to\infty} \mathbb{E}\|\mathbf{X}^\star - \hat{\mathbf{X}}_{\text{opt}}\|^2/d = \rho - q_x.$

Theorem (informal) : If either

a) $\Phi_{\mu i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_\beta(0,1)$ (standard Gaussian distribution)

b) P_0 is Gaussian and $\boldsymbol{\Phi} = \mathbf{W}\mathbf{B}$

Gaussian matrix   Any matrix

, the replica conjecture stands.

- We use probabilistic adaptive interpolation methods [Barbier&al '19], based on the seminal works of [Guerra '03, Talagrand '06].
- The replica formula for non-linear GLMs was so far only proven for real Gaussian matrices [Barbier&al '19], we tackle for the first time heavily correlated data !

Algorithmic limits

Strong conjecture : The optimal polynomial-time algorithm is an explicit iterative algorithm :

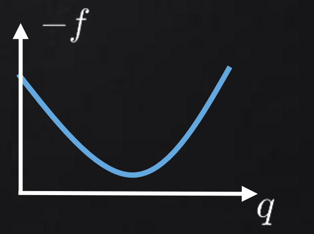
Approximate Message Passing, called here G-VAMP *(Generalized Vector Approximate Message Passing).*

[Mézard '89, Donoho&al '09, Montanari&al '10 , Krzakala&al '11, Rangan&al '16, Schniter&al '16, ...]

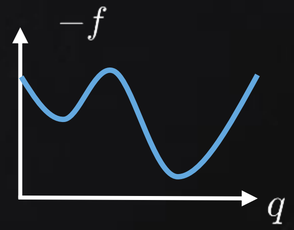
$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\ln \mathcal{Z}_d(\mathbf{Y},\mathbf{\Phi}) = \sup_{q_x, q_z} \underbrace{[I_0^{(\beta)}(q_x) + I_{\mathrm{out}}^{(\beta)}(q_z) + \beta I_{\mathrm{int}}(q_x, q_z)]}_{\text{"Replica-symmetric" potential } f(q_x,q_z)}$$

Important result [Schniter&al '16] : The MSE of G-VAMP in the large n limit is given by a *fixed-point algorithm* on the replica-symmetric potential starting from $q_x = q_z = 0$ (random initialization).

⟹ We can investigate "computational-to-statistical" gaps by studying the landscape of $f(q_x, q_z)$!



No gap: AMP-easy



Gap : AMP-hard

Application: thresholds in phase retrieval

$P_{\text{out}}(y|z) = P_{\text{out}}(y||z|)$

Weak-recovery

What is the minimal number of measurements $\alpha = n/d$ necessary to beat a random guess in polynomial time ?

This threshold $\alpha_{\text{WR,Algo}}$ is the only solution to :

$$\alpha = \frac{\langle\lambda\rangle_\nu^2}{\langle\lambda^2\rangle_\nu}\left[1 + \left\{\int_{\mathbb{R}} \mathrm{d}y \frac{\left(\int_{\mathbb{K}} \mathcal{D}_\beta z\ (|z|^2 - 1)\ P_{\text{out}}\left[y\middle|\sqrt{\frac{\rho\langle\lambda\rangle_\nu}{\alpha}}z\right]\right)^2}{\int_{\mathbb{K}} \mathcal{D}_\beta z\ P_{\text{out}}\left[y\middle|\sqrt{\frac{\rho\langle\lambda\rangle_\nu}{\alpha}}z\right]}\right\}^{-1}\right]$$

For any phase/sign retrieval channel, the highest weak recovery threshold is reached by random column-orthogonal/unitary matrices.

Derived by a stability analysis of the replica-symmetric potential around the uninformative point.

Strong recovery

Noiseless phase retrieval $P_{\text{out}}(y|z) = \delta(y - |z|^2)$ and Gaussian prior

How many measurements $\alpha = n/d$ are necessary to be able to information-theoretically achieve the best possible recovery ?

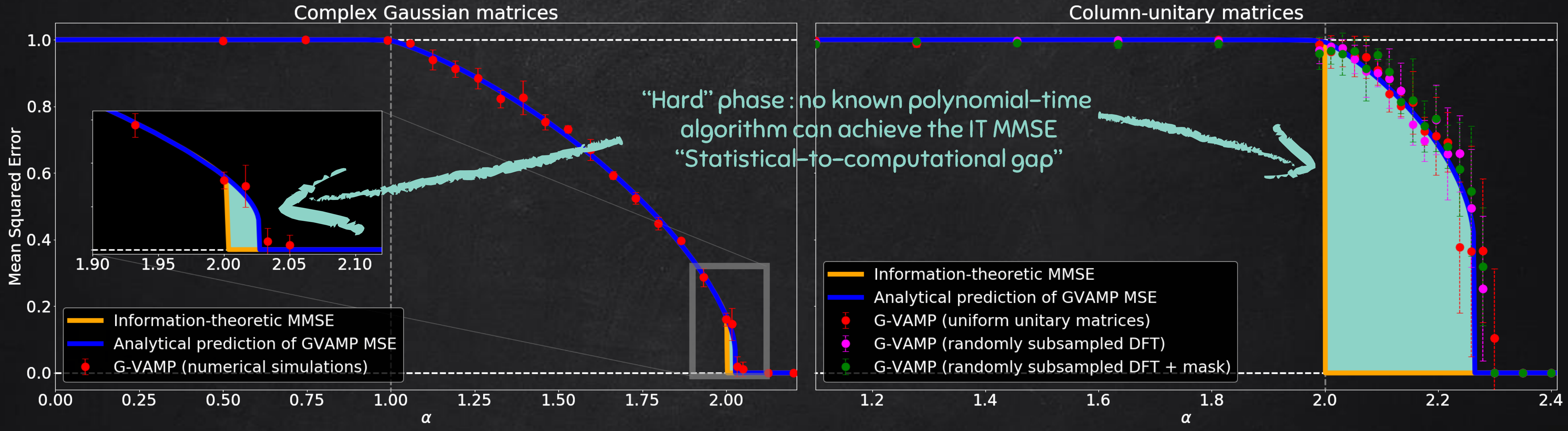
If (a.s.) $\frac{1}{d}\mathrm{rk}\left(\frac{\mathbf{\Phi}^\dagger\mathbf{\Phi}}{d}\right) \to r \in [0,1]$ then $\alpha_{\text{FR,IT}} = \beta r$

Analysis of the global maxima of the replica-symmetric potential.

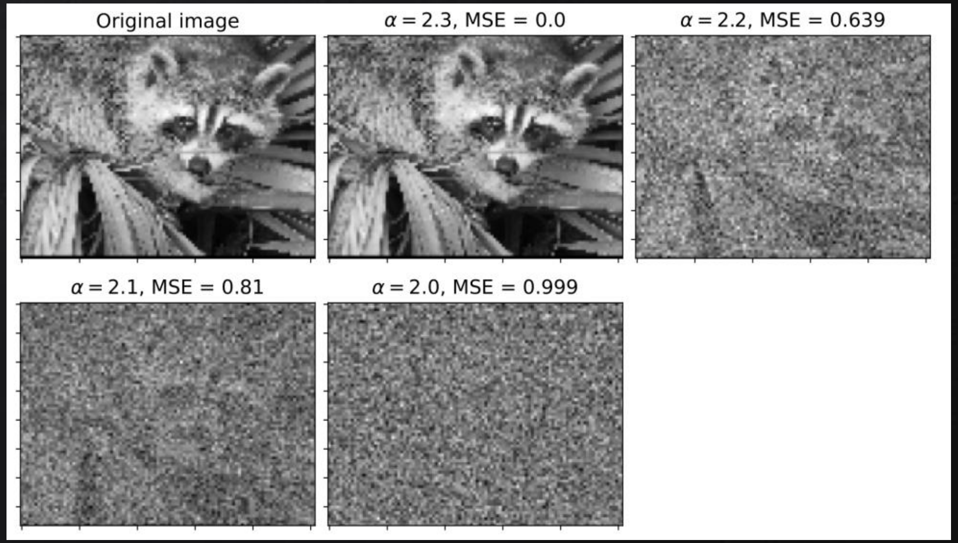
- The real case $\alpha_{\text{FR,IT}} = r$ can be derived by a counting argument. [Candès&Tao '05]
- The complex case $\alpha_{\text{FR,IT}} = 2r$ can (as far as we know) only be derived our analysis of the replica-symmetric potential !

Numerical applications

$P_{\text{out}}(y|z) = \delta(y - |z|^2)$ and a Gaussian prior



- G-VAMP matches the analytical predictions, <u>even with a natural image</u> !
- Matrices with controlled structure (e.g. randomly subsampled DFT) still perform very well !
- For column-unitary matrices we have $\alpha_{\text{FR,IT}} = \alpha_{\text{WR,Algo}} = 2$: "all-or-nothing" IT transition.
- For all other full-rank complex matrices $\alpha_{\text{WR,Algo}} < \alpha_{\text{FR,IT}}$.



Cheaper algorithms?

- SDP relaxations [Candès&al '15a&b, Waldspurger&al '15, Goldstein&al '18, ...]
- Non-convex optimization procedures [Netrapalli&al '15, Candès&al '15c, ...]
- Approximate Message-Passing (this talk) [Barbier&al '19, A.M.&al '20...]

Computationally heavy / Need informed initialization → Spectral methods

[Mondelli&al '18, Luo&al '18, Dudeja&al '19,...]

Given a phase retrieval problem with an arbitrary sensing matrix, we want an "optimal" spectral method in terms of MSE:

$$\boxed{\mathrm{MSE} \equiv \frac{1}{d\rho} \|\mathbf{X}^* - \hat{\mathbf{X}}_{\mathrm{spectral}}\|^2}$$

This talk: Two different strategies, related to the statistical physics approach to high-dimensional inference.

- Method I : Linearization of message-passing algorithms.
- Method II : Bethe Hessian analysis from the Thouless-Anderson-Palmer [TAP77] free energy.

Optimal spectral method

$$\partial_\omega g_{\rm out}(y_\mu, 0, \sigma^2) = -\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\frac{\int_{\mathbb{K}} dx\, e^{-\frac{\beta}{2\sigma^2}|x|^2}\,|x|^2\, P_{\rm out}(y_\mu|x)}{\int_{\mathbb{K}} dx\, e^{-\frac{\beta}{2\sigma^2}|x|^2}\, P_{\rm out}(y_\mu|x)}$$

Main conjecture: The optimal spectral method (in terms of achieved error) in the class $\mathrm{M}(\mathcal{T}) \equiv \frac{1}{d}\sum_{\mu=1}^{n} \mathcal{T}(y_\mu)\boldsymbol{\Phi}_\mu \boldsymbol{\Phi}_\mu^\dagger$ is $\mathrm{M}(\mathcal{T}^\star)$:

$$\boxed{\mathcal{T}^*(y) = \frac{\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)}{1 + \frac{\rho\langle\lambda\rangle_\nu}{\alpha}\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)}}$$

From Method II: "$\mathrm{M}_{\rm TAP}$"

- In noiseless phase retrieval one has $\mathcal{T}^\star(y) = 1 - 1/y$.
- Fully constructive derivation of the optimal method: we did not assume the method to be in the class $\mathrm{M}(\mathcal{T})$!
- The optimal spectral method does not depend on the spectrum of the sensing matrix (apart from a global scaling) !

➡ Consequences for practitioners: one only needs to know the observation channel to construct the method!

Our other approach (Method I) gives a slightly different result:

Linearized Approximate Message-Passing (LAMP) spectral method.

$$\mathrm{M}_{\rm LAMP} \equiv \frac{\rho\langle\lambda\rangle_\nu}{\alpha}\left(\frac{\alpha}{\langle\lambda\rangle_\nu}\frac{\boldsymbol{\Phi}\boldsymbol{\Phi}^\dagger}{d} - \mathrm{I}_n\right)\mathrm{Diag}(\{\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)\}) \Longrightarrow \hat{\mathbf{x}} \equiv \frac{\boldsymbol{\Phi}^\dagger \mathrm{Diag}(\{\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)\})\hat{\mathbf{u}}}{\|\boldsymbol{\Phi}^\dagger \mathrm{Diag}(\{\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)\})\hat{\mathbf{u}}\|}\sqrt{d\rho}.$$

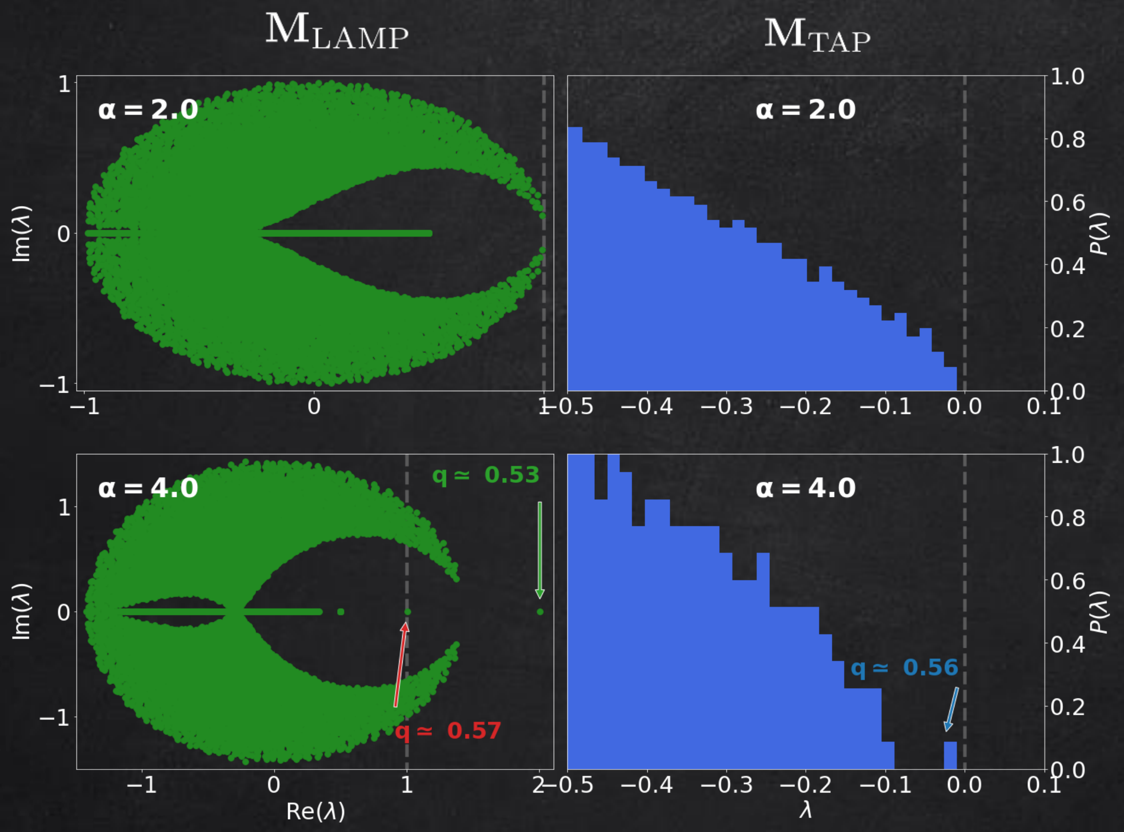
$\mathrm{M}_{\rm LAMP}$ is a $n \times n$ non-Hermitian matrix (complex spectrum).

$\hat{\mathbf{u}}$: top eigenvector of $\mathrm{M}_{\rm LAMP}$.

Comparing spectral methods

Complex Gaussian Φ and Poisson noise

$$P_{\text{out}}(y|z) = e^{-\Lambda|z|^2} \sum_{k=0}^{\infty} \delta(y-k) \frac{\Lambda^k |z|^{2k}}{k!}$$



$$\Lambda = 1$$

$$q = \frac{1}{d} \sum_{i=1}^{d} X_i^* \hat{x}_i$$

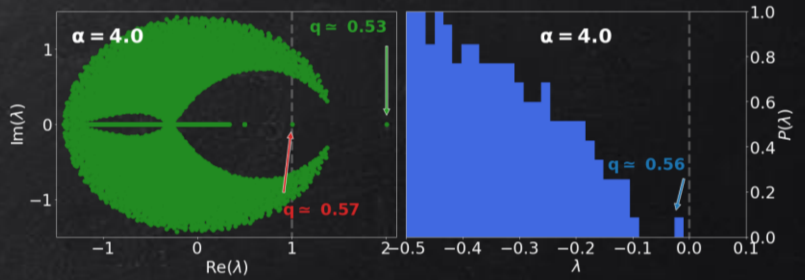
- The optimal method corresponds to marginal stability in both $\mathrm{M_{LAMP}}$ and $\mathrm{M_{TAP}}$.
- But message-passing algorithms and the TAP approach are fundamentally equivalent ! [A.M.&al '19]

Puzzle: why is the dominant eigenvector of $\mathrm{M_{LAMP}}$ a suboptimal estimator ?

"Marginality vs instability" puzzle

Similar phenomenon noticed in community detection [Dall'Amico&al ' 19, 21].

Spectral methods performance



Noiseless complex phase retrieval $Y_\mu = \frac{1}{d}|\Phi \mathbf{X}^*|^2$

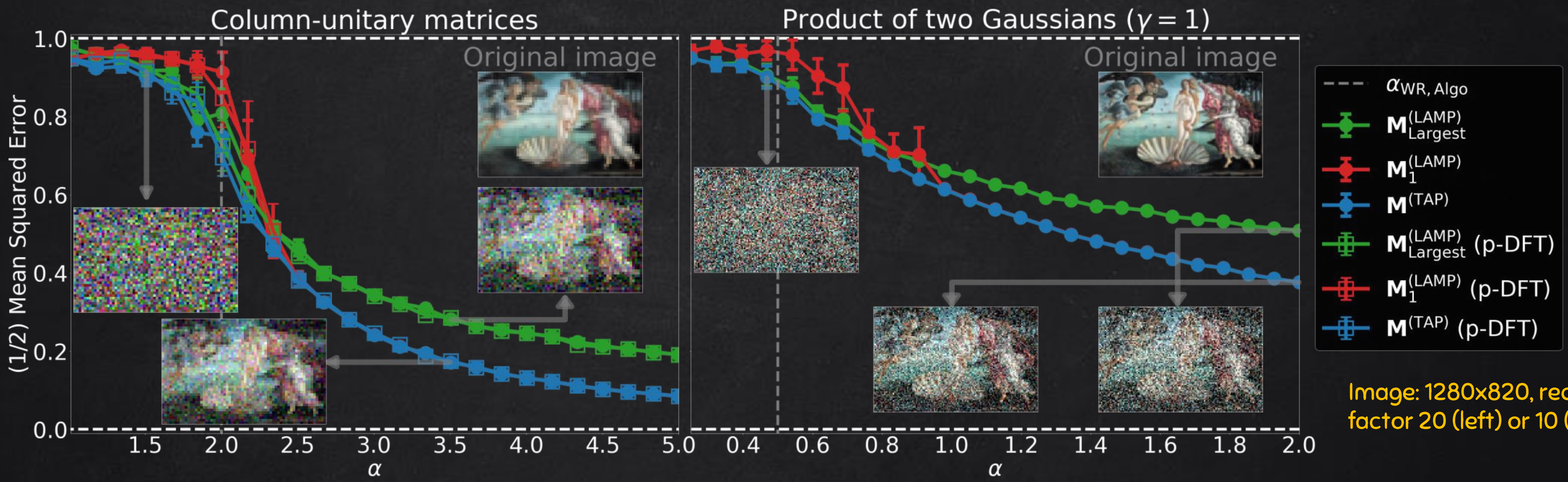


Image: 1280x820, reduced by a factor 20 (left) or 10 (right).

- $\hat{x}_{\text{LAMP}}(\lambda = 1) \Longleftrightarrow \hat{x}_{\text{TAP}}$, achieving the best overlap. Otherwise $\hat{x}_{\text{LAMP}}(\lambda_{\max})$ is suboptimal in terms of MSE.
- Our theory stays valid for matrices with controlled structure (partial DFT $\equiv$ randomly subsampled DFT).
- For partial DFT matrices, we use the method as initialization of a gradient-descent procedure: perfect recovery at $\alpha \in (3, 4)$, while the best polynomial-time algorithm achieves $\alpha_{\text{PR}} \simeq 2, 3$. ➡ Very competitive while computationally cheap!

Conclusion

(new results in red)

Matrix ensemble and value of β	$\alpha_{\mathrm{WR,Algo}}$	$\alpha_{\mathrm{FR,IT}}$	$\alpha_{\mathrm{FR,Algo}}$
Real Gaussian $\mathbf{\Phi}$ $(\beta = 1)$	0.5	1	$\simeq 1.12$
Complex Gaussian $\mathbf{\Phi}$ $(\beta = 2)$	1	2	$\simeq 2.027$
Real column-orthogonal $\mathbf{\Phi}$ $(\beta = 1)$	1.5	1	$\simeq 1.584$
Complex column-unitary $\mathbf{\Phi}$ $(\beta = 2)$	2	2	$\simeq 2.265$
$\mathbf{\Phi} = \mathbf{W}_1\mathbf{W}_2$ $(\beta = 1$, aspect ratio $\gamma)$	$\gamma/(2(1+\gamma))$	$\min(1,\gamma)$	Theorem
$\mathbf{\Phi} = \mathbf{W}_1\mathbf{W}_2$ $(\beta = 2$, aspect ratio $\gamma)$	$\gamma/(1+\gamma)$	$\min(2, 2\gamma)$	Theorem
$\mathbf{\Phi}$, $\beta \in \{1,2\}$, $\mathrm{rk}[\mathbf{\Phi}^\dagger\mathbf{\Phi}]/n = r$	Analytical expression	βr	Conjecture
Gauss. $\mathbf{\Phi}$, $\beta \in \{1,2\}$, symm. P_0, P_{out}	Analytical expression	Theorem	Theorem
$\mathbf{\Phi} = \mathbf{W}\mathbf{B}$, $\beta \in \{1,2\}$, Gauss. P_0, symm. P_{out}	Analytical expression	Theorem	Theorem
$\mathbf{\Phi}$, $\beta \in \{1,2\}$, symm. P_0, P_{out}	Analytical expression	Conjecture	Conjecture

Noiseless phase retrieval with Gaussian prior

Generic phase retrieval with any prior

I

Fundamental limits of phase retrieval

II

Spectral methods

- Constructive derivation of a conjecturally optimal spectral method in generic phase retrieval problems.
- Our results apply to randomly subsampled DFT matrices and to real image recovery.

"Marginality vs instability" puzzle: Bethe Hessian and message-passing constructions of spectral methods should be equivalent!

THANK YOU !