Construction of optimal spectral methods in phase retrieval

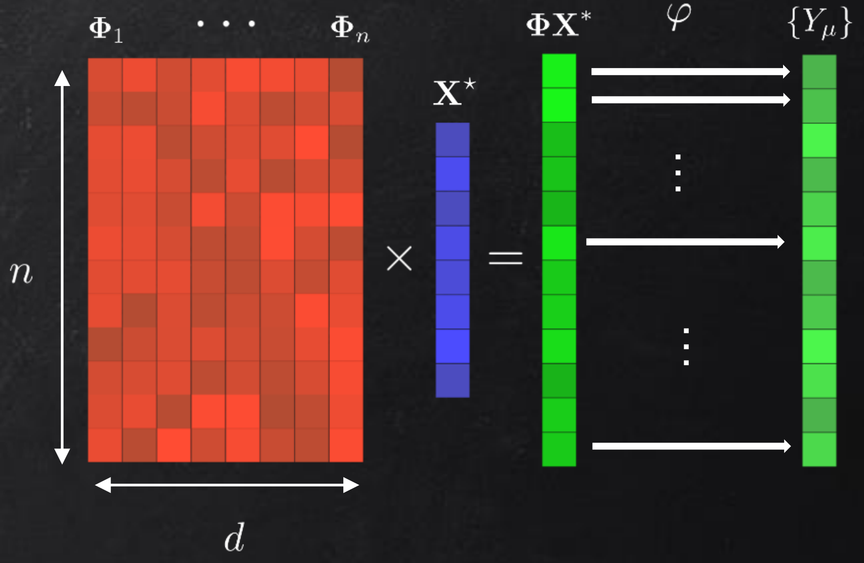
arXiv:2012.04524 ; *MSML'21.*

Antoine Maillard

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

Random Matrix Theory and Networks – June 16th 2021

Inference in high dimensions



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

Generalized Linear Model (GLM)

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

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

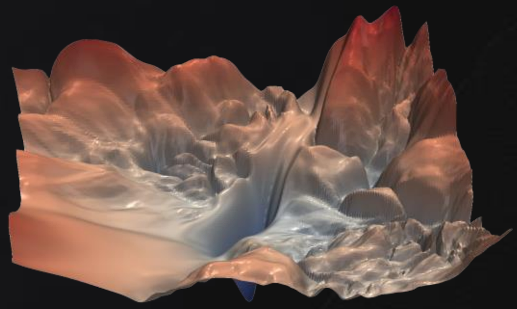
(Probabilistic) channel with possible noise.

Sensing matrix (real/complex)

Signal (real/complex), d-dimensional

Real / Complex
$\beta = 1$ $\beta = 2$

In general, one needs to solve a highly non-convex optimization problem in high dimensions.



Example (empirical risk minimization – square loss): $\hat{\mathbf{X}} = \mathrm{argmin}_{\mathbf{x}} \sum_{\mu=1}^n \left(Y_\mu - \varphi(\mathbf{x} \cdot \mathbf{\Phi}_\mu / \sqrt{n})\right)^2$

Image credits: [Li&al '18]

Phase retrieval

$$Y_\mu \sim P_{\rm out}\left(\cdot \middle| \frac{1}{\sqrt{d}} \sum_{i=1}^d \Phi_{\mu i} X_i^\star\right)$$

In phase retrieval, one measures the modulus $P_{\rm out}(y|z) = P_{\rm out}(y||z|)$, e.g. noiseless $Y_\mu = \frac{1}{d}|(\mathbf{\Phi}\mathbf{X}^\star)_\mu|^2$; Poisson-noise $Y_\mu \sim \mathrm{Pois}(\Lambda|(\mathbf{\Phi}\mathbf{X}^\star)_\mu|^2/d)$.

- Classical problem of learning with a non-convex landscape.
- Arises in signal processing, statistical estimation, optics, X-ray crystallography, astronomy, microscopy… : optical detectors lose information on the phase of the signals.

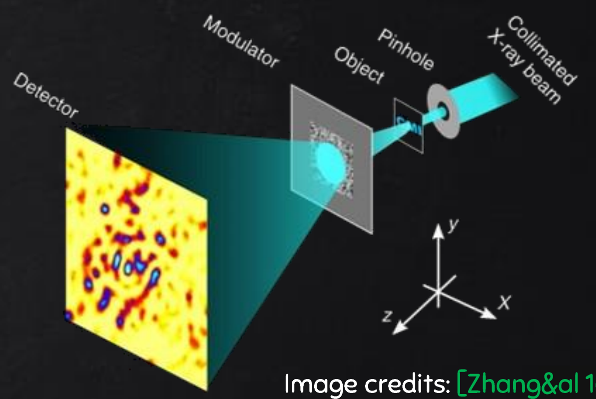


Image credits: [Zhang&al 16]

How to solve this problem efficiently in high dimensions ? $n, d \to \infty$

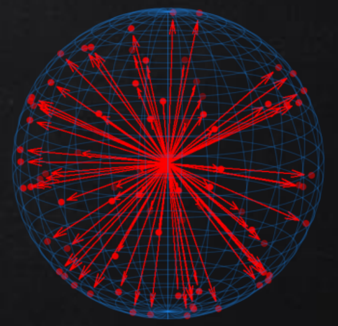
- SDP relaxations [Candès&al '15a&b, Waldspurger&al '15, Goldstein&al '18, …]
- Non-convex optimization procedures [Netrapalli&al '15, Candès&al '15c, …]
- Approximate Message-Passing [Barbier&al '19, A.M.&al '20]

Computationally heavy / Need informed initialization → Spectral methods

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

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

The bulk of eigenvalues of $\Phi^\dagger\Phi/d$ 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 $S_i^2 \overset{\text{i.i.d.}}{\sim} \nu$.

Weak-recovery threshold in phase retrieval [A.M.&al '20]

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

➡ If the signal $\mathbf{X}^\star$ has norm $\frac{1}{n}\|\mathbf{X}^*\|^2 = \rho$, this threshold $\alpha_{\text{WR,Algo}}$ is the only solution to :

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

This is an implicit equation.

Example: Noiseless phase retrieval : $\boxed{\alpha = \left(1 + \frac{\beta}{2}\right)\frac{\langle\lambda\rangle_\nu^2}{\langle\lambda^2\rangle_\nu}}$

- Gaussian matrices : $\alpha_{\text{WR,Algo}} = \frac{\beta}{2}$ [Barbier&al '19, Mondelli &al '19, Luo&al '19]
- Random column-orthogonal/unitary matrices : $\alpha_{\text{WR,Algo}} = 1 + \frac{\beta}{2}$ [Dudeja&al '20] for $\beta = 2$

Construction of spectral methods

Given a generic phase retrieval problem, we want to design spectral methods such that:

- Weak-recovery is achieved for all $\alpha > \alpha_{\mathrm{WR,Algo}}$.
- For all α the method is optimal among all possible spectral methods in terms of estimation error:

d

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

<u>This talk:</u> Three different strategies, related to the statistical-physics approach to high-dimensional inference.

- <u>Method I :</u> Naïve generalization of what is known for Gaussian matrices.
- <u>Method II :</u> Linearization of message-passing algorithms.
- <u>Method III :</u> Bethe Hessian analysis from the Thouless-Anderson-Palmer [TAP77] free energy.

Method I: naive approach

$$y_\mu \sim P_{\rm out}\Big(\cdot \Big| \frac{1}{\sqrt{d}} \sum_{i=1}^d \Phi_{\mu i} X_i^*\Big)$$

Most previous works reduced to methods of the type $\boxed{\mathbf{M}(\mathcal{T}) \equiv \frac{1}{d}\sum_{\mu=1}^n \mathcal{T}(y_\mu)\mathbf{\Phi}_\mu \mathbf{\Phi}_\mu^\dagger}$

[Chen&al '15, Wang&al '16, Zhang&al '17, Mondelli&al '19, Luo&al '19, Dudeja&al '19]

Idea: For large sample size $n \gg d$ we expect $\mathbf{M} \simeq \mathbb{E}[\mathbf{M}] = a\mathrm{I}_n + b\mathbf{X}^*(\mathbf{X}^*)^\dagger$.

For Gaussian matrices $\mathbf{\Phi}$ the optimal method in this class is given by $\boxed{\mathcal{T}^\star_{\rm Gaussian}(y) \equiv \frac{\partial_\omega g_{\rm out}(y_\mu, 0, \rho)}{1 + \rho \partial_\omega g_{\rm out}(y_\mu, 0, \rho)}}$

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

$(\mathbb{K} = \mathbb{R}, \mathbb{C})$

- In noiseless phase retrieval one has $\mathcal{T}^\star_{\rm Gaussian}(y) = 1 - 1/y$.
- It achieves weak recovery at the optimal threshold $\alpha = \alpha_{\rm WR, Algo}$.
- Optimal also in terms of achieved Mean Squared Error among the class of $\mathrm{M}(\mathcal{T})$.
- We can naively use it for all matrices: $\boxed{\mathrm{M}_{\rm naive} \equiv \mathrm{M}(\mathcal{T}^\star_{\rm Gaussian})}$.

Method II: Linearization of Message-Passing

- [Schniter&al '16, A.M.&al '20]: For GLMs with rotationally-invariant matrices, the best-known polynomial-time algorithm (in terms of estimation error) is given by *Generalized Vector Approximate Message-Passing* (G-VAMP).
- But G-VAMP is computationally very expensive $\longrightarrow$ Construct a spectral method from the G-VAMP iterations.

Symmetry of the phase retrieval problem $P_{\text{out}}(y|z) = P_{\text{out}}(y||z|)$ $\longrightarrow$ G-VAMP has a trivial fixed point at $\hat{\mathbf{x}} = 0$.

Linearize G-VAMP around this point.

Linearized Approximate Message-Passing (LAMP) spectral method.

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

M_{LAMP} is a $n \times n$ non-Hermitian matrix (complex spectrum).

$\hat{\mathbf{u}}$: top eigenvector of M_{LAMP}.

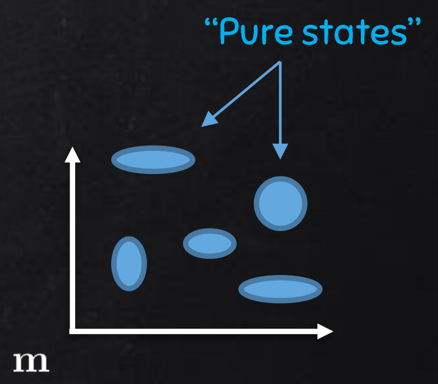
Similar approaches have been applied in community detection [Krzakala&al '13], phase retrieval with unitary matrices [Ma&al '21] and spiked matrix estimation [Aubin&al '20].

Method III: TAP landscape and Bethe Hessian

Thouless-Anderson-Palmer approach [TAP77]

- The posterior measure of $\mathbf{x}|\mathbf{Y}$ (the *Gibbs measure*) decomposes along pure states.
- These pure states can be found by "tilting" the measure, imposing $m_i = \langle x_i \rangle$ and $\sigma_i^2 = \mathrm{Var}(x_i)$:

  They are the maxima of the free entropy of this constrained measure, as a function of $(\mathbf{m}, \boldsymbol{\sigma})$.



- TAP free entropy for rotationally-invariant generalized linear models derived in [A.M.&al '19], generalizing [Parisi&Potters '95]:

Involved but explicit!

$$f_{\mathrm{TAP}}(\mathbf{m}) = \sup_{\sigma \geq 0} \sup_{\substack{\mathbf{g} \in \mathbb{K}^n \\ r \geq 0}} \mathrm{extr}_{\substack{\boldsymbol{\omega} \in \mathbb{K}^n \\ b \geq 0}} \mathrm{extr}_{\substack{\boldsymbol{\lambda} \in \mathbb{K}^d \\ \gamma \geq 0}} \Big[\frac{\beta}{d} \sum_{i=1}^{d} \lambda_i \cdot m_i + \frac{\beta\gamma}{2d} \big(d\sigma^2 + \sum_{i=1}^{d} |m_i|^2\big) - \frac{\beta}{d} \sum_{\mu=1}^{n} \omega_\mu \cdot g_\mu - \frac{\beta b}{2d} \big(\sum_{\mu=1}^{n} |g_\mu|^2 - \alpha d r\big) + \frac{1}{d} \sum_{i=1}^{d} \ln \int_{\mathbb{K}} P_0(\mathrm{d}x) e^{-\frac{\beta\gamma}{2}|x|^2 - \beta\lambda_i \cdot x}$$

$$+ \frac{\alpha}{n} \sum_{\mu=1}^{n} \ln \int_{\mathbb{K}} \frac{\mathrm{d}h}{\left(\frac{2\pi b}{\beta}\right)^{\beta/2}} P_{\mathrm{out}}(y_\mu|h) e^{-\frac{\beta|h-\omega_\mu|^2}{2b}} + \frac{\beta}{d} \sum_{i=1}^{d} \sum_{\mu=1}^{n} g_\mu \cdot \big(\frac{\Phi_{\mu i}}{\sqrt{d}} m_i\big) + \beta F(\sigma^2, r) \Big].$$

$$F(x,y) \equiv \inf_{\zeta_x, \zeta_y > 0} \Big[\frac{\zeta_x x}{2} + \frac{\alpha \zeta_y y}{2} - \frac{\alpha - 1}{2} \ln \zeta_y - \frac{1}{2} \langle \ln(\zeta_x \zeta_y + \lambda) \rangle_\nu \Big] - \frac{1}{2} \ln x - \frac{\alpha}{2} \ln y - \frac{1+\alpha}{2}.$$

<u>Weak-recovery impossible</u> $\alpha < \alpha_{\mathrm{WR}}$

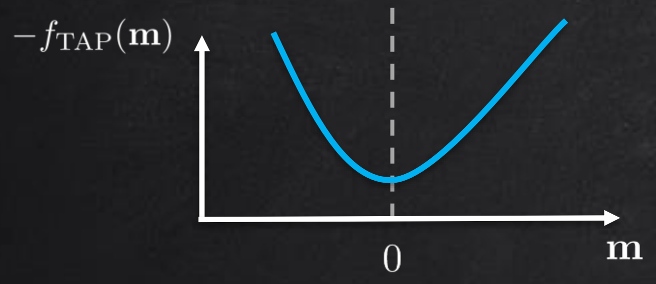
Global maximum of f_{TAP} in $\mathbf{m} = 0$: the uninformative "paramagnetic" point.

<u>Weak-recovery possible</u> $\alpha > \alpha_{\mathrm{WR}}$

$\mathbf{m} = 0$ is an unstable stationary point of f_{TAP}, which has a global maximum in $\mathbf{m} \neq 0$ (optimal estimator).

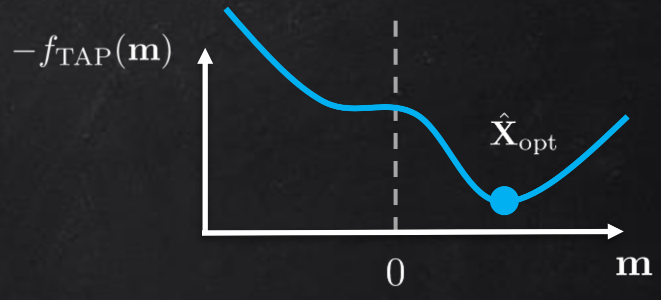
Weak-recovery impossible $\alpha < \alpha_{\rm WR}$

Global maximum of $f_{\rm TAP}$ in $\mathbf{m} = 0$: the uninformative "paramagnetic" point.



Weak-recovery possible $\alpha > \alpha_{\rm WR}$

$\mathbf{m} = 0$ is an unstable stationary point of $f_{\rm TAP}$, which has a global maximum in $\mathbf{m} \neq 0$ (optimal estimator) .



A spectral method can only use the physical information available in the uninformative point $\mathbf{m} = 0$.

Compute the Hessian of $f_{\rm TAP}$ at the paramagnetic point.

Constructive derivation of a spectral method that is conjectured to be optimal.

TAP – Bethe Hessian spectral method.

$$\mathbf{M}_{\rm TAP} \equiv -d\nabla^2 f_{\rm TAP}(\mathbf{m}=0) = -\frac{1}{\rho}\mathrm{I}_d + \frac{1}{d}\sum_{\mu=1}^{n} \frac{\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)}{1 + \frac{\rho\langle\lambda\rangle_\nu}{\alpha}\partial_\omega g_{\rm out}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)} \mathbf{\Phi}_\mu \mathbf{\Phi}_\mu^\dagger$$

Similar to previous strategies in community detection.
[Saade&al '14]

Relations between the methods and the marginality puzzle

Analytical study of the three spectral methods

- $\mathrm{M}_{\mathrm{TAP}}$ is the "naïve" generalization of Method I.
- $\mathrm{M}_{\mathrm{TAP}}$ and $\mathrm{M}_{\mathrm{LAMP}}$: transition at the optimal $\alpha_{\mathrm{WR,Algo}}$.

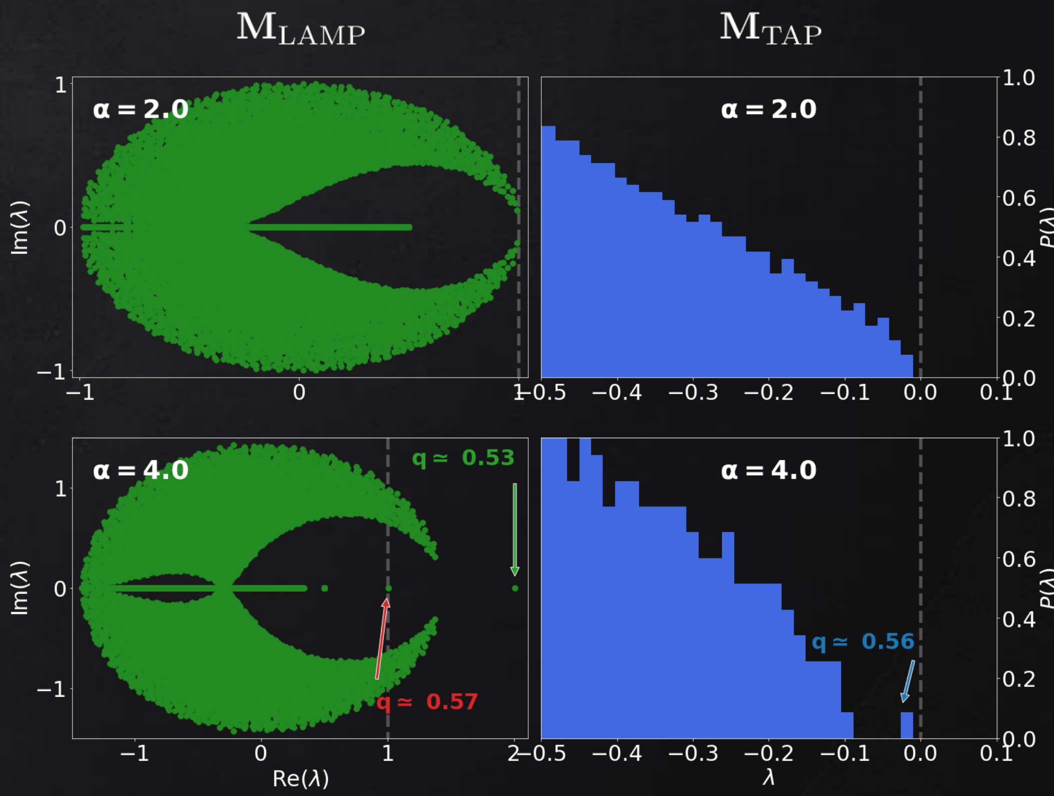
- The fixed points of the TAP free entropy are in exact correspondence with the ones of G-VAMP [A.M.&al '19].
- Marginal stability: When recovery is possible, the largest eigenvalue of $\mathrm{M}_{\mathrm{TAP}}$ concentrates on 0, and is in exact correspondence with an eigenvalue of $\mathrm{M}_{\mathrm{LAMP}}$ that concentrates on 1.
- Instability of $\mathrm{M}_{\mathrm{LAMP}}$: No other eigenvector of $\mathrm{M}_{\mathrm{TAP}}$ achieves non-trivial performance, while the dominant eigenvalue of $\mathrm{M}_{\mathrm{LAMP}}$ is another non-trivial estimator that is suboptimal.

Puzzling disparity between methods that should be equivalent.

- Complex Gaussian Φ
- Poisson channel ($\Lambda = 1$)

$$P_{\mathrm{out}}(y|z) = e^{-\Lambda|z|^2} \sum_{k=0}^{\infty} \delta(y-k) \frac{\Lambda^k |z|^{2k}}{k!} \Longrightarrow \alpha_{\mathrm{WR,Algo}} = 2$$

- We denote the overlap $q = \frac{1}{d} \sum_{i=1}^{d} X_i^* \hat{x}_i$



Optimal spectral method

$$\mathrm{M}(\mathcal{T}) \equiv \frac{1}{d}\sum_{\mu=1}^{n} \mathcal{T}(y_\mu)\mathbf{\Phi}_\mu\mathbf{\Phi}_\mu^\dagger$$

From the Bethe Hessian analysis

Main conjecture: For any right-orthogonally invariant sensing matrix, the optimal spectral method (in terms of weak-recovery threshold and achieved error) belongs to the class of matrices $\mathrm{M}(\mathcal{T})$ and is attained in:

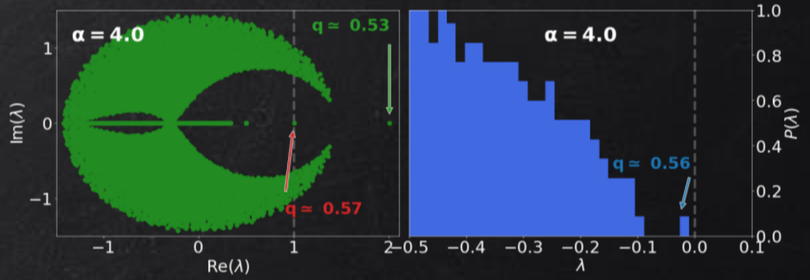
$$\boxed{\mathcal{T}^*(y) = \frac{\partial_\omega g_{\mathrm{out}}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)}{1 + \frac{\rho\langle\lambda\rangle_\nu}{\alpha}\partial_\omega g_{\mathrm{out}}(y_\mu, 0, \rho\langle\lambda\rangle_\nu/\alpha)}}$$

- We did not assume anything on the form of the method, yet the optimal spectral method we constructed is in the class of $\mathrm{M}(\mathcal{T})$ matrices: we confirm the validity of the restriction of previous works on spectral methods!
- The optimal spectral method does not depend on the spectrum of the sensing matrix (apart from a global scaling), nor on the sampling ratio α!

➡ Very different from the optimal algorithms ! [A.M.&al '20]

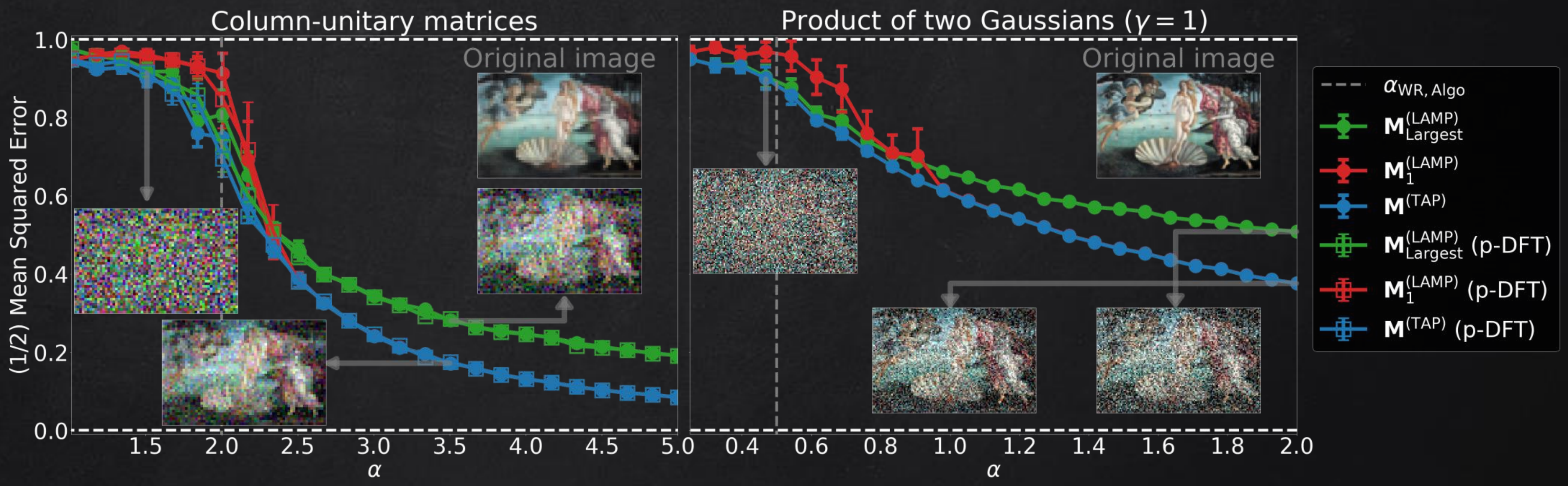
➡ Consequences for practitioners: one only needs to know the observation channel to construct the method!

Spectral methods performance



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

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



- All three methods share the same weak-recovery threshold, in agreement with the best polynomial-time algorithm.
- $\hat{x}_{\text{LAMP}}(\lambda = 1) \Longleftrightarrow \hat{x}_{\text{TAP}}$, 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)

Spectral initialization in noiseless phase retrieval

We combine the optimal spectral method $\mathrm{M}_{\mathrm{TAP}}$ with gradient descent on the square loss $L(\mathbf{x}) \equiv \frac{1}{2n}\sum_{\mu=1}^{n}\left\{\left|\frac{(\Phi\mathbf{x})_\mu}{\sqrt{d}}\right|^2 - \left|\frac{(\Phi\mathbf{X}^\star)_\mu}{\sqrt{d}}\right|^2\right\}^2.$

- Noiseless phase retrieval with randomly subsampled DFT sensing matrix. [A.M &al '20]
- Already perfect recovery at $\alpha \in (3,4)$. For partial-DFT, perfect recovery with the best polynomial-time algorithm is $\alpha_{\mathrm{PR}} \simeq 2,3$

⟹ Very competitive while computationally cheap!

Conclusion and perspectives

Main contributions

- Constructive derivation of a conjecturally optimal spectral method in generic phase retrieval problems, in a framework that encompasses real/complex variables and a wide variety of sensing matrices.
- Our results apply to randomly subsampled DFT matrices and to real image (i.e. structured signal) recovery.
- We use two fundamentally equivalent approaches – message-passing linearization and Bethe Hessian analysis – that yield the same optimal performance, associated with a marginal stability of the linear dynamics.

Theory far from complete

- The "marginality vs instability" puzzle: In $\mathrm{M}_{\mathrm{LAMP}}$ the optimal method is "hidden" inside the bulk and marginally stable, while the dominant eigenvalue is unstable and suboptimal.
- What if we do not know how the data was generated ? What become of the thresholds and of the performance of the spectral estimators ? → A first observation: marginality disappears when using a mismatched channel distribution.

THANK YOU !