A few challenges in large-rank matrix denoising and factorization

Antoine Maillard

Vittorio Erba, Florent Krzakala, Marc Mézard,

Emanuele Troiani, Lenka Zdeborová

Journal of Statistical Mechanics 2022

Mathematical and Scientific Machine Learning 2022

ETH zürich

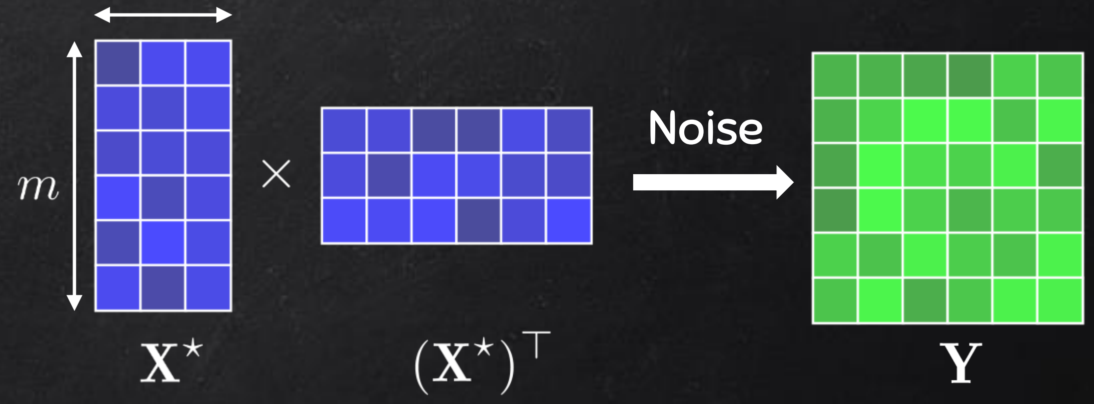
21st INFORMS Applied Probability Society Conference, Nancy– June 30th 2023

Setting

[M., Krzakala, Mézard & Zdeborová '22]

$$Y_{\mu\nu} \sim P_{\text{out}}\Big(\cdot \,\Big|\, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X^\star_{\mu i} X^\star_{\nu i}\Big)$$

$\mathbf{X}^\star \sim P_X$



Task

Recover $\mathbf{X}^\star$ — Factorization

Recover $\mathbf{S}^\star = \mathbf{X}^\star(\mathbf{X}^\star)^\top$ — Denoising

Dictionary learning, sparse coding, sparse PCA, matrix completion...

Setting:

- High-dimensional: $m \to \infty$
- P_X, P_{out} are known ➡ $\hat{\mathbf{S}}(\mathbf{Y}) = \mathbb{E}[\mathbf{S}|\mathbf{Y}]$

Minimal Mean Squared Error estimator

$\mathrm{MMSE} = \mathbb{E}\|\mathbb{E}[\mathbf{S}|\mathbf{Y}] - \mathbf{S}^\star\|_F^2$

- Large m limit ?
- Reachable by efficient algorithms ?

✓ $n = \mathcal{O}(1)$ (low-rank).

[Rangan&al'12], [Deshpande&al'14], [Lesieur&al '15], [Perry&al'16], [Lelarge&al'19], [Aubin, M.,&al '20] ....

Here $m/n = \alpha \in (0, \infty)$

This talk

Denoising ➕ $X_{\mu i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Generalizations

Other models for $\mathbf{S}$, non-symmetric factorization $\mathbf{Y} \sim P_{\text{out}}(\mathbf{UV})$, ...

Rotationally-invariant denoising (1)

[Ledoit & Péché '11]
[Bun, Allez, Bouchaud & Potters '16]

Assume noise is additive and rotationally-invariant

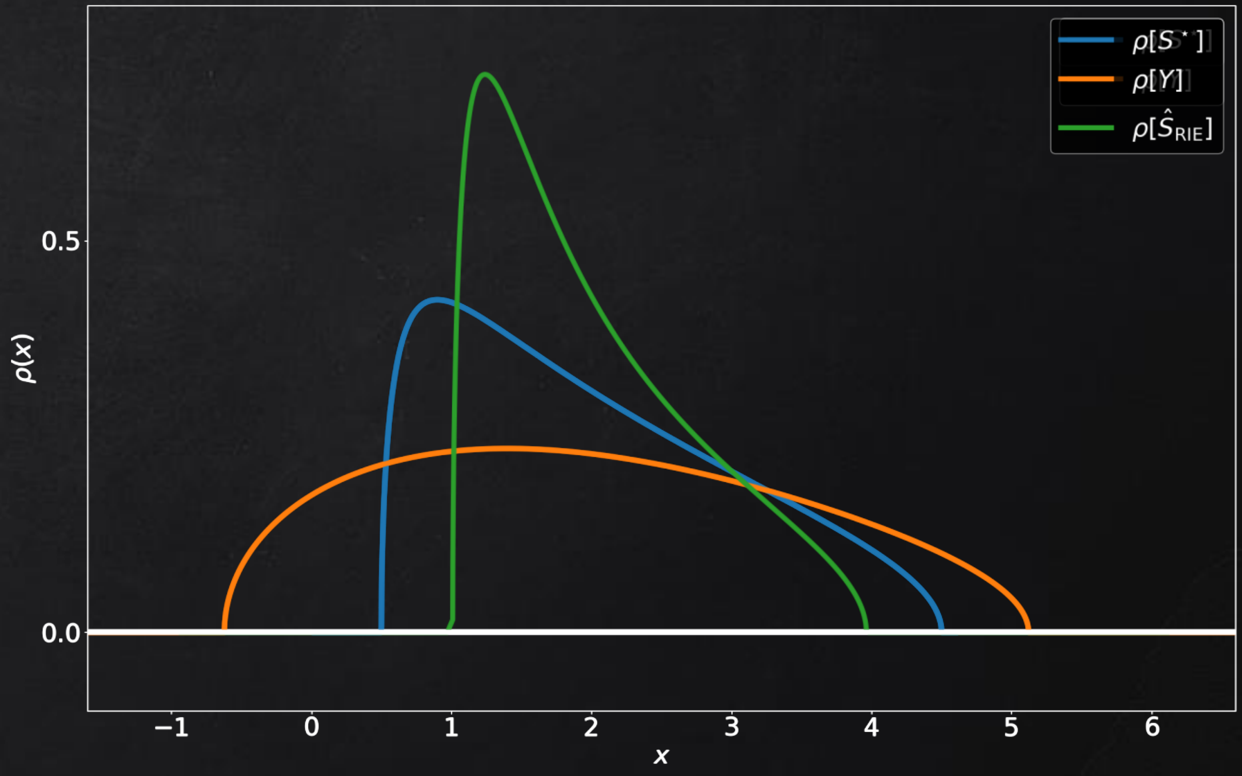
$$\boxed{\mathbf{Y} = \mathbf{S}^\star + \sqrt{\Delta}\mathbf{Z}}$$ → Gaussian (GOE) noise

Rotationally-invariant estimator (RIE) $\hat{\mathbf{S}}_{\rm RIE} = \sum_{\mu=1}^{m} \hat{\xi}_\mu \mathbf{u}_\mu \mathbf{u}_\mu^\top$ → Eigenvectors of $\mathbf{Y}$

$$\hat{\xi}_\mu = \arg\min_{\boldsymbol{\xi}\in\mathbb{R}^m} \|\mathbf{S}^\star - \hat{\mathbf{S}}(\boldsymbol{\xi})\|_F^2$$

Eigenvalues of $\mathbf{Y}$

$$\boxed{\hat{\xi}_\mu = y_\mu - 2\Delta \times \mathrm{P.P.}\int \rho_{\mathbf{Y}}(\mathrm{d}t)\frac{1}{t - y_\mu}}$$



"Miraculous solution": only depends on the spectral properties of $\mathbf{Y}$

Easily computed with tools of free probability [Voiculescu, ...]

Rotationally-invariant denoising (2)

$$\mathbf{Y} = \mathbf{S}^\star + \sqrt{\Delta}\mathbf{Z}$$

$$\mathbb{P}(\mathbf{S}|\mathbf{Y}) = \frac{1}{\mathcal{Z}(\mathbf{Y})} P^{(\alpha)}_{\mathrm{Wish.}}(\mathbf{S}) \exp\left\{-\frac{1}{4\Delta}\|\mathbf{Y}-\mathbf{S}\|_F^2\right\} \propto P^{(\alpha)}_{\mathrm{Wish.}}(\mathbf{S}) \exp\left\{-\frac{\|\mathbf{S}\|_F^2}{4\Delta} + \frac{1}{2\Delta}\mathrm{Tr}[\mathbf{YS}]\right\}$$

$$f(\mathbf{ODO}^\top) = f(\mathbf{D}) \Rightarrow \mathbb{E}[f(\mathbf{S})|\mathbf{Y}] = \int_{\mathbb{R}^m} \mathrm{d}\mathbf{D}\, q(\mathbf{D})\, f(\mathbf{D}) \int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O} \exp\left\{\frac{1}{2\Delta}\mathrm{Tr}\left[\mathbf{YODO}^\top\right]\right\}$$

"HCIZ" integral [Harish-Chandra-Itzykson-Zuber]

$$\boxed{\int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O} \exp\left\{m\mathrm{Tr}\left[\mathbf{AO}\mathbf{B}\mathbf{O}^\top\right]\right\}}$$

Full-rank

known for $m \to \infty$

- When $\mathrm{rank}(\mathbf{A}) = o(m)$ [Guionnet '05]
- When $\mathrm{rank}(\mathbf{A}) = \Theta(m)$ [Matytsin '94, Guionnet&al '02]

[M., Krzakala, Mézard & Zdeborová '22]

- Analytical formula $\mathrm{MMSE} = \int \rho_{\mathbf{Y}}(\mathrm{d}t)\,(\cdots)$ cf also [Pourkamali, Barbier & Macris '23]
- Re-derivation of the optimal RIE estimator as $\hat{\mathbf{S}}_{\mathrm{opt.}} = \mathbb{E}[\mathbf{S}|\mathbf{Y}] \simeq \hat{\mathbf{S}}_{\mathrm{RIE}}$

Estimator and asymptotic characterization extend to the non-symmetric setting [Erba, Troiani, Krzakala, M. & Zdeborová '22]

Beyond rotation invariant denoising

[M., Krzakala, Mézard & Zdeborova '22]

$$Y_{\mu\nu} \sim P_{\rm out}(\cdot|S^\star_{\mu\nu})$$

$$\mathbf{S}^\star = \frac{1}{\sqrt{n}}\mathbf{X}^\star(\mathbf{X}^\star)^\top$$

- Exact characterization of the MMSE ✗
- Efficient optimal estimator ✗

Unknown

Proposed Approximate Message Passing (AMP) algorithms

→ Convergence problems + hard-to-control assumptions

[Kabashima & al '16, Parker &al '14, Zou & al '21, Lucibello & al '21]

This talk: sketch a perturbative approach to clarify these difficulties, and lay a path for improvement.

$$S_{\mu\nu} = \frac{1}{\sqrt{n}}\sum_{i=1}^n x_{\mu i}x_{\nu i}$$

$H_{\mu\nu}$: conjugate field

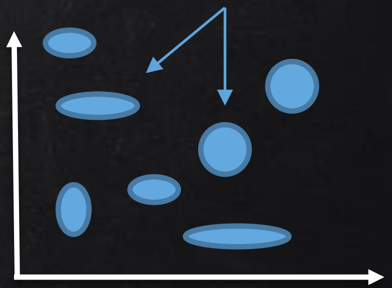
$$\mathbb{P}(\mathbf{X}|\mathbf{Y}) = \frac{1}{\mathcal{Z}(\mathbf{Y})}\prod_{\mu,i} p_X(x_{\mu i})\prod_{\mu,\nu} P_{\rm out}\left(Y_{\mu\nu}\middle|\frac{1}{\sqrt{n}}\sum_{i=1}^n x_{\mu i}x_{\nu i}\right)$$

$$P(\mathbf{X},\mathbf{H}) \propto \left[\prod_{\mu,i} p_X(x_{\mu i})\right]\left[\prod_{\mu,\nu} Q(Y_{\mu\nu},H_{\mu\nu})\right]\exp\left\{\frac{1}{\sqrt{n}}\sum_{i,\mu,\nu} H_{\mu\nu}x_{\mu i}x_{\nu i}\right\}$$

"Effective" distribution

Plefka–Georges–Yedidia expansion

"Pure states"

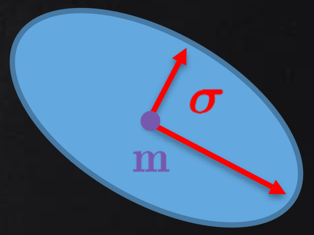


Idea: $P(\mathbf{X},\mathbf{H}) \propto \left[\prod_{\mu,i} p_X(x_{\mu i})\right]\left[\prod_{\mu,\nu} Q(Y_{\mu\nu},H_{\mu\nu})\right] \exp\left\{\frac{\eta}{\sqrt{n}}\sum_{i,\mu,\nu} H_{\mu\nu}x_{\mu i}x_{\nu i}\right\}$

Thouless-Anderson-Palmer approximation [TAP77]

There is a function $\Phi_{\mathrm{TAP}}(\mathbf{m},\boldsymbol{\sigma})$ whose maxima give the "pure states" in which $P(\mathbf{X},\mathbf{H})$ concentrates its mass.

TAP free energy

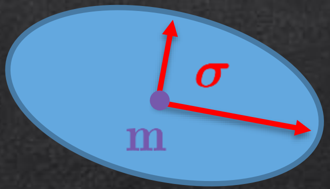


Worked out in spin glass models and simpler statistical inference models [Parisi&Potters '95], [M.&al '19]

$$\Phi_{\mathrm{TAP}}(\mathbf{m},\boldsymbol{\sigma}) = \sum_{k=0}^{\infty} \frac{\partial_\eta^k \Phi_{\mathrm{TAP}}(\mathbf{m},\boldsymbol{\sigma})[\eta=0]}{k!}$$

- $\partial_k^\eta \Phi_{\mathrm{TAP}}(\mathbf{m},\boldsymbol{\sigma})[\eta=0]$ can be recursively computed by the "PGY" method [Plefka '82, Georges&Yedidia '91]
- It turns out that (at least for the first orders) "$\eta \Leftrightarrow \sqrt{m/n} = \sqrt{\alpha}$": "overcomplete" limit $\mathbf{S}^\star \simeq \mathrm{I}_m + \varepsilon(\alpha)$.

The PGY expansion



$$\mathbf{m} = (m_{\mu\nu})$$

$$\boldsymbol{\sigma} = \mathrm{Diag}(\{\sigma_{\mu\nu}\})$$

PGY

$$\Phi_{\mathrm{TAP}}(\mathbf{m},\boldsymbol{\sigma}) = \sum_{\mu,\nu} \Big[\operatorname*{extr}_{\boldsymbol{\omega},\mathbf{b}} \Big\{ -\omega_{\mu\nu} m_{\mu\nu} - \frac{b_{\mu\nu}}{2}\big(-\sigma_{\mu\nu}^2 + m_{\mu\nu}^2\big) + \ln \int \mathrm{d}z \, \frac{e^{-\frac{1}{2b_{\mu\nu}}(z-\omega_{\mu\nu})^2}}{\sqrt{2\pi b_{\mu\nu}}} P_{\mathrm{out}}(Y_{\mu\nu}|z) \Big\} \Big]$$

$$+\frac{\eta^2}{2}\sum_{\mu,\nu}[m_{\mu\nu}^2 - \sigma_{\mu\nu}^2] + \frac{\eta^3}{6n^{1/2}} \sum_{\substack{\mu_1,\mu_2,\mu_3 \\ \text{pairwise distinct}}} m_{\mu_1\mu_2} m_{\mu_2\mu_3} m_{\mu_3\mu_1} + \mathcal{O}(\eta^4)$$

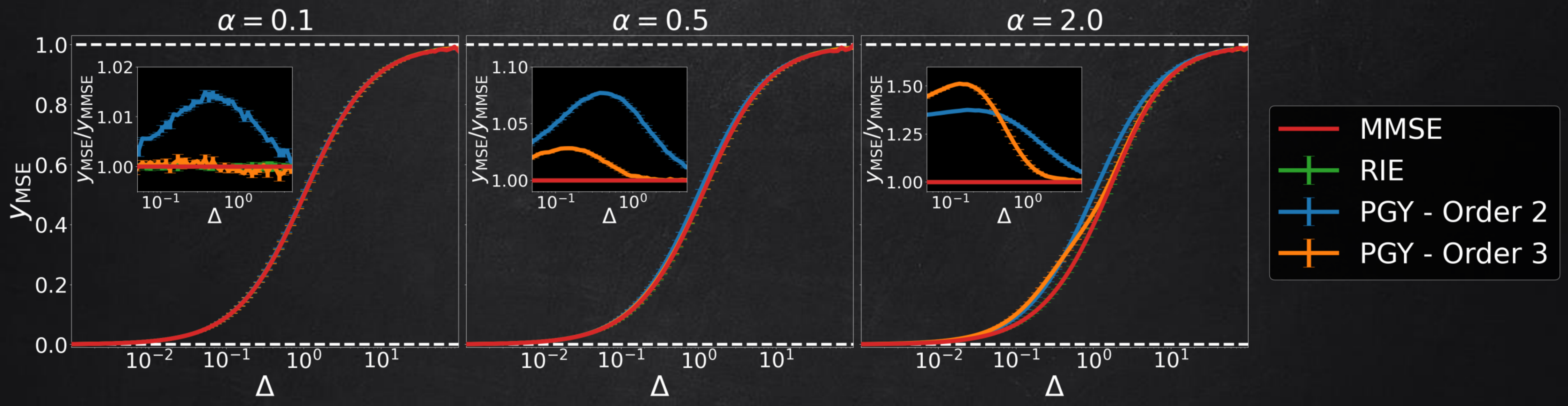
- Iterative equations to find the maxima of Φ_{TAP} can be turned into an algorithm [M., Foini, &al '19]
- Truncating at order η^2 $\Longleftrightarrow$ "AMP" algorithms of [Kabashima & al'16, …]

  → We explicit their approximation
- However, order η^3 and above are <u>not negligible</u>

  → [Kabashima & al'16, …] effectively neglect some 3rd order correlations

Numerics for Gaussian denoising

$$\mathbf{Y} = \frac{\mathbf{X}^\star (\mathbf{X}^\star)^\intercal}{\sqrt{n}} + \sqrt{\Delta}\mathbf{Z} \quad ; \quad m = \alpha n$$



- "PGY order 3" significantly improves over order 2, in the overcomplete regime $\alpha \ll 1$.
- Analytical check that $\hat{\mathbf{S}}_{\mathrm{PGY}} \simeq \mathbb{E}[\mathbf{S}|\mathbf{Y}]$ up to order $(\sqrt{\alpha})^3$ ✓

Limitation of the PGY method

Orders 1, 2, 3, … of the expansion ⇒ Educated conjecture about arbitrary orders

For orders ≥ 4, PGY expansion becomes very tedious, need more investigation !

Conclusion

Some (of the many) open directions

- PGY expansion at orders ≥ 4 ? Arbitrary orders ? Possible resummation of the series ?
- Efficient denoising/factorization algorithms when $n = \Theta(m)$ and for non-RI noise ? $Y_{\mu\nu} \sim P_{\rm out}(\cdot|\sqrt{m}S^{\star}_{\mu\nu})$
- Transition between low-rank and extensive-rank regimes when rotationally-invariant :

$$I(\mathbf{A}) = \frac{1}{m^2} \log \int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O} \exp\left\{m \mathrm{Tr}\left[\mathbf{AOBO}^\top\right]\right\} \qquad \mathrm{rank}(A) = o(m) \longleftrightarrow \mathrm{rank}(A) = \Theta(m)$$

?

[Alice Guionnet, "Rare events in Random Matrix theory", ICM 2022]

Other recent works: [Camilli & Mézard '23, Barbier & Macris '23, Pourkamali & Macris '23, Landau, Mel & Ganguli '23]

THANK YOU !