Some advances on extensive-rank matrix factorization & denoising

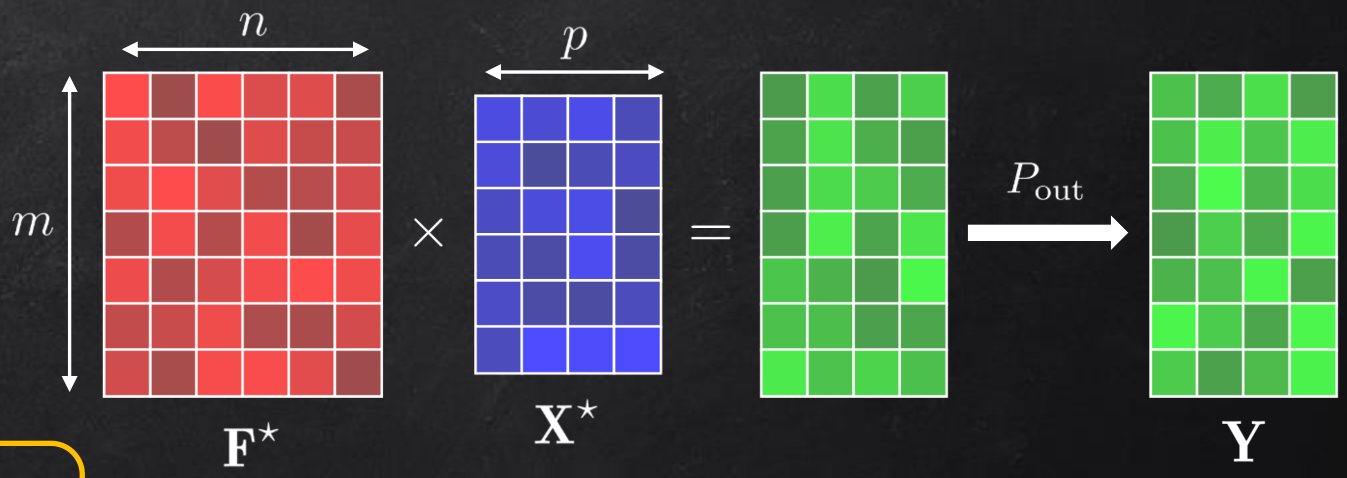
arXiv:2110.08775

Antoine Maillard, Florent Krzakala, Marc Mézard, Lenka Zdeborová

EPFL – November 1st 2021

Matrix factorization

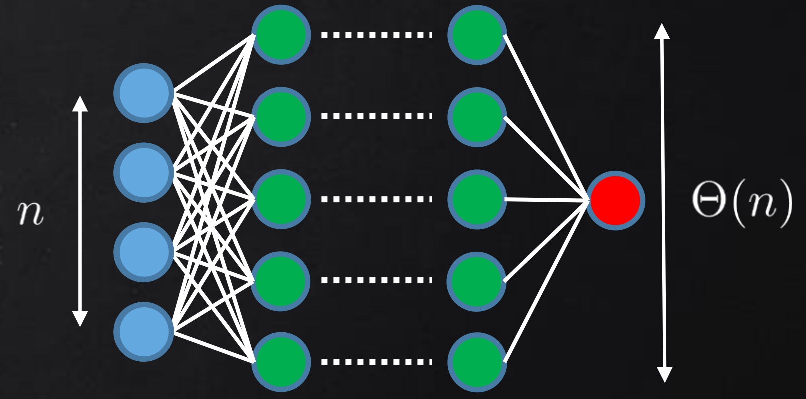
$$Y_{\mu l} \sim P_{\text{out}}\left(\cdot \,\Big|\, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F^\star_{\mu i} X^\star_{il} \right)$$



Prior information: $F^\star_{\mu i} \overset{\text{i.i.d.}}{\sim} P_F$ and $X^\star_{il} \overset{\text{i.i.d.}}{\sim} P_X$

- High-dimensional limit: $n, m, p \to \infty$
- Bayes-optimal setting: priors and channel are known.

Wide and high-dimensional neural nets



The "classical" stat-phys toolbox

- ~~Replica method~~
- ~~Cavity method / Message-passing~~

Previous attemps failed !
[Kabashima & al '16]

Some recent progress on replicas for Gaussian channels [Barbier&Macris '21]

A closely related problem: symmetric matrix factorization

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

$X^\star_{\mu i} \overset{\text{i.i.d.}}{\sim} P_X$

$m/n \to \alpha > 0$

Very different from low-rank

$$\mathbf{Y} = \frac{1}{\sqrt{m}} \mathbf{x}\mathbf{x}^\intercal + \sqrt{\Delta}\mathbf{Z}$$

The Plefka-Georges-Yedidia expansion

[Plefka '82, Georges&Yedidia '91]

Original Gibbs measure

$$\Phi_\beta = \frac{1}{n} \ln \sum_{\{S_i = \pm 1\}} e^{-\beta H(\{S_i\})}$$

Constraint $\langle S_i \rangle = m_i$

TAP free entropy

$$\Phi_\beta(\{m_i\}) = \operatorname*{extr}_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^n \lambda_i m_i + \frac{1}{n} \ln \sum_{\{S_i\}} e^{-\beta H(\{S_i\}) - \sum_i \lambda_i S_i} \right\}$$

High-temperature (low-β) expansion

- Initiated by Plefka, made much more general and systematic by Georges and Yedidia.

- In finite-rank problems: yields correct TAP equations in a wide range of rotationally-invariant models.
- Equivalent to the message-passing approach.

[A.M.&al '19]

Goal: Apply the PGY formalism to the posterior distribution of symmetric matrix factorization

$$\Phi_{\mathbf{Y},n} \equiv \frac{1}{nm} \ln \int \prod_{\mu,i} P_X(\mathrm{d}X_{\mu i}) \prod_{\mu<\nu} P_{\mathrm{out}}\Big(Y_{\mu\nu} \Big| \frac{1}{\sqrt{n}} \sum_i X_{\mu i} X_{\nu i}\Big)$$

PGY expansion in symmetric matrix factorization

Step 1: Write $\Phi_{\mathbf{Y},n}$ in a suitable form for PGY expansion.

Fourier transform of the delta

$$e^{nm\Phi_{\mathbf{Y},n}} = \int \prod_{\mu,i} P(\mathrm{d}X_{\mu i}) \prod_{\mu<\nu} \left[\int \mathrm{d}\hat{H}_{\mu\nu} P_{\mathrm{out}}(Y_{\mu\nu}|\hat{H}_{\mu\nu}) \delta\Big(\hat{H}_{\mu\nu} - \frac{1}{\sqrt{n}} \sum_i X_{\mu i} X_{\nu i}\Big) \right] = \int \prod_{\mu<\nu} P^{\mu\nu}_{H,Y}(\mathrm{d}H_{\mu\nu}) \prod_{\mu,i} P_X(\mathrm{d}X_{\mu i}) e^{-H_{\mathrm{eff}}[\mathbf{X},\mathbf{H}]}$$

Prior distribution of the conjugate field

$$\boxed{P^{\mu\nu}_{H,Y}[\mathrm{d}H] \equiv \int \frac{\mathrm{d}\hat{H}}{2\pi} e^{iH\hat{H}} P_{\mathrm{out}}(Y_{\mu\nu}|\hat{H})}$$

Effective interaction Hamiltonian

$$\boxed{H_{\mathrm{eff}}[\mathbf{X},\mathbf{H}] \equiv \frac{1}{\sqrt{n}} \sum_{\mu<\nu} \sum_i (iH)_{\mu\nu} X_{\mu i} X_{\nu i}}$$

Step 2: $e^{-H_{\mathrm{eff}}} \to e^{-\eta H_{\mathrm{eff}}}$ $\qquad \eta$: "Inverse temperature"

Small – η expansion of the TAP free entropy, fixing the first and second moments:

$$\langle X_{\mu i} \rangle = m_{\mu i} \qquad \langle (iH)_{\mu\nu} \rangle = -g_{\mu\nu}$$

$$\langle X^2_{\mu i} \rangle = v_{\mu i} + (m_{\mu i})^2 \qquad \langle (iH)^2_{\mu\nu} \rangle = -r_{\mu\nu} + g^2_{\mu\nu}$$

First orders of the series

$$nm\Phi_{\mathbf{Y},n}(0) = \sum_{\mu,i}\Big[\lambda_{\mu i}m_{\mu i} + \frac{\gamma_{\mu i}}{2}\big(v_{\mu i} + (m_{\mu i})^2\big) + \ln\int P_X(\mathrm{d}x)\,e^{-\frac{\gamma_{\mu i}}{2}x^2 - \lambda_{\mu i}x}\Big] + \sum_{\mu<\nu}\Big[-\omega_{\mu\nu}g_{\mu\nu} - \frac{b_{\mu\nu}}{2}\big(-r_{\mu\nu} + g_{\mu\nu}^2\big) + \ln\int \mathrm{d}z\,\frac{e^{-\frac{1}{2b_{\mu\nu}}(z-\omega_{\mu\nu})^2}}{\sqrt{2\pi b_{\mu\nu}}}P_{\mathrm{out}}(Y_{\mu\nu}|z)\Big].$$

$$nm[\Phi_{\mathbf{Y},n}(\eta) - \Phi_{\mathbf{Y},n}(0)] = \frac{\eta}{\sqrt{n}}\sum_{\substack{\mu<\nu\\ i}} g_{\mu\nu}m_{\mu i}m_{\nu i} - \frac{\eta^2}{2n}\sum_{\substack{\mu<\nu\\ i}} r_{\mu\nu}[v_{\mu i}v_{\nu i} + v_{\mu i}m_{\nu i}^2 + m_{\mu i}^2 v_{\nu i}] + \frac{\eta^2}{4n}\sum_{\mu,\nu,i} g_{\mu\nu}^2 v_{\mu i}v_{\nu i} + \frac{\eta^3}{6n^{3/2}}\sum_i \sum_{\substack{\mu_1,\mu_2,\mu_3\\ \text{pairwise distinct}}} g_{\mu_1\mu_2}g_{\mu_2\mu_3}g_{\mu_3\mu_1}v_{\mu_1 i}v_{\mu_2 i}v_{\mu_3 i} + \mathcal{O}(\eta^4).$$

One should extremize with respect to all parameters $(\mathbf{m}, \mathbf{v}, \mathbf{g}, \mathbf{r}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{b})$ → TAP equations

- TAP equations of $\Phi_{\mathbf{Y},n}$ at order η^2 ⟺ Fixed point of the GAMP algorithm of [Kabashima&al'16] I

- But order 3 is not negligible !

II $$\frac{1}{n}\sum_{\substack{\mu_1,\mu_2,\mu_3\\ \text{pairwise distinct}}} \frac{g_{\mu_1\mu_2}}{\sqrt{n}}\frac{g_{\mu_2\mu_3}}{\sqrt{n}}\frac{g_{\mu_3\mu_1}}{\sqrt{n}} \underset{n\to\infty}{\simeq} c_3\Big(\frac{\mathbf{g}}{\sqrt{n}}\Big)$$ Free cumulant [A.M.&al '19]

I + II → Clear evidence of where the approximation of [Kabashima&al'16] fails.

Going to higher orders ?

Intrinsic limitation of the PGY method

Compute order 1, 2, 3, ... of the expansion → Educated conjecture about arbitrary orders

Order 3: $\frac{1}{3!}\left(\frac{\partial^3 \Phi_{\mathbf{Y},n}}{\partial \eta^3}\right)_{\eta=0} = \frac{1}{6n^{5/2}m}\sum_i \sum_{\substack{\mu_1,\mu_2,\mu_3 \\ \text{pairwise distinct}}} g_{\mu_1\mu_2} g_{\mu_2\mu_3} g_{\mu_3\mu_1} \prod_{a=1}^{3} v_{\mu_a i} + o_n(1)$

Free cumulant $c_3(\mathbf{g}/\sqrt{n})$

Similar to finite-rank problems [A.M.&al'19] → Possible conjecture $\frac{\partial^k \Phi_{\mathbf{Y},n}}{\partial \eta^k} \propto c_k(\mathbf{g}/\sqrt{n})$

Seems to lead to inconsistencies...

The matrix $\mathbf{g}$ is a parameter of the problem → Spectrum of $\mathbf{g}$ = order parameter ?

Difference from finite-rank problems

Conclusion

Still largely open problem, we are investigating !

Matrix Denoising: a simplified problem

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

Denoising problem

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

Only interested in the recovery of $\mathbf{S}^\star$

Many possible choices:

- Wishart matrix: $\mathbf{S}^\star = \mathbf{X}^\star (\mathbf{X}^\star)^\intercal / \sqrt{nm}$
- Wigner (GOE) matrix: $S^\star_{\mu\nu} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- Uniformly-sampled symmetric orthogonal matrix: $\mathbf{S}^\star = \mathbf{O}$
- ...

- Free entropy: $\exp\{nm\Phi_{\mathbf{Y},n}\} = \int P_S(\mathrm{d}\mathbf{S}) \frac{e^{-\frac{1}{4\Delta}\sum_{\mu,\nu}\left(Y_{\mu\nu} - \sqrt{m} S_{\mu\nu}\right)^2}}{(2\pi\Delta)^{\frac{m(m-1)}{4}}}$

For $\mathbf{S}^\star$ a Wishart matrix:

$$\Phi^{\text{factorization}}_{\mathbf{Y},n} = \Phi^{\text{denoising}}_{\mathbf{Y},n}$$

- If $P_{\text{out}}(Y|\cdot) = \mathcal{N}(Y, \Delta)$, the optimal rotationally-invariant estimator is studied in [Bun&al '16].

Idea: Use denoising as a controlled setup to probe our PGY expansion for factorization.

The PGY expansion for denoising

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

$$\boxed{\begin{aligned} nm\Phi_{\mathbf{Y},n} &= \sum_{\mu<\nu} \Big[-\omega_{\mu\nu} g_{\mu\nu} - \frac{b_{\mu\nu}}{2}\big(-r_{\mu\nu} + g_{\mu\nu}^2\big) + \ln \int \mathrm{d}z \, \frac{e^{-\frac{1}{2b_{\mu\nu}}(z-\omega_{\mu\nu})^2}}{\sqrt{2\pi b_{\mu\nu}}} P_{\rm out}(Y_{\mu\nu}|z) \Big] \\ &+ \frac{\eta^2}{2n} \sum_{\mu<\nu} [g_{\mu\nu}^2 - r_{\mu\nu}] + \frac{\eta^3}{6mn^{5/2}} \sum_{\substack{\mu_1,\mu_2,\mu_3 \\ \text{pairwise distinct}}} g_{\mu_1\mu_2} g_{\mu_2\mu_3} g_{\mu_3\mu_1} + \mathcal{O}(\eta^4) \end{aligned}}$$

Similar to factorization, but we do not fix $(\mathrm{m}, \boldsymbol{\sigma})$.

<u>Meaning of g:</u> $g_{\mu\nu} = \Big\langle \frac{\partial_{\hat{H}} P_{\rm out}(Y_{\mu\nu}|\hat{H}_{\mu\nu})}{P_{\rm out}(Y_{\mu\nu}|\hat{H}_{\mu\nu})} \Big\rangle_{\mathbf{Y}}$ $\xrightarrow[P_{\rm out}(Y|\cdot) = \mathcal{N}(Y, \Delta)]{\text{Gaussian channel}}$ $g_{\mu\nu} = \frac{1}{\Delta}\Big[Y_{\mu\nu} - \frac{1}{\sqrt{n}} \Big\langle \sum_{i=1}^n X_{\mu i} X_{\nu i} \Big\rangle \Big]$

Denoising estimator

<u>Remarks:</u>
- The PGY expansion is actually an expansion in powers of $\alpha = m/n$!
- For a Gaussian channel, g is <u>diagonal in the eigenbasis of $\mathbf{Y}$</u>!

$$\left. \begin{aligned} \mathbf{Y}/\sqrt{m} &= \sum_y y \, \mathbf{v}_y \mathbf{v}_y^\intercal \\ \mathbf{g}/\sqrt{m} &= \sum_y g_y \mathbf{v}_y \mathbf{v}_y^\intercal \end{aligned} \right\} \xrightarrow[\mathbb{E} X_{\mu i}^2 = 1]{} \boxed{g_y = \frac{y}{\Delta+1} + \frac{(\Delta + 1 - y^2)}{(\Delta+1)^3}\sqrt{\alpha} + \mathcal{O}(\alpha)}$$

"PGY order-3" denoiser

"PGY order-2"

The free entropy

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

For factorization and denoising

$$\exp\{nm\Phi_{\mathbf{Y},n}\} = \int P_S(\mathrm{d}\mathbf{S}) \frac{e^{-\frac{1}{4\Delta}\sum_{\mu,\nu}\left(Y_{\mu\nu}-\sqrt{m}S_{\mu\nu}\right)^2}}{(2\pi\Delta)^{\frac{m(m-1)}{4}}}$$

Wishart distribution

$$P_S(\mathrm{d}\mathbf{S}) \propto (\det \mathbf{S})^{\frac{n-m-1}{2}} e^{-\frac{\sqrt{nm}}{2}\mathrm{Tr}\,\mathbf{S}}\mathrm{d}\mathbf{S}$$

$$\mathbf{S} = \mathbf{OLO}^\intercal$$

$$\boxed{\exp\{nm\Phi_{\mathbf{Y},n}\} = C(n,m)\int_{\mathbb{R}_+^m} \mathrm{d}\mathbf{L} \prod_{\mu<\nu} |l_\mu - l_\nu| e^{-\frac{m}{2}\sum_{\mu=1}^m\left(\frac{l_\mu}{\sqrt{\alpha}}+\frac{l_\mu^2}{2\Delta}\right)+\frac{n-m-1}{2}\sum_{\mu=1}^m \ln l_\mu}\int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O}\, e^{\frac{\sqrt{m}}{2\Delta}\mathrm{Tr}[\mathbf{YOLO}^\intercal]}}$$

Extensive-rank HCIZ integral

$$\boxed{\int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O}\, e^{\frac{\sqrt{m}}{2\Delta}\mathrm{Tr}[\mathbf{YOLO}^\intercal]} \simeq \exp\left\{\frac{m^2}{2} I_\Delta[\rho_{\mathbf{Y}},\rho_{\mathbf{S}}]\right\}}$$

[Matytsin '94, Guionnet&al '02]

$$\boxed{\begin{aligned}\Phi_{\mathbf{Y},n} = C(\alpha) + \sup_{\rho_{\mathbf{S}}}\Bigg\{&\frac{\alpha}{4}\int \rho_{\mathbf{S}}(\mathrm{d}x)\rho_{\mathbf{S}}(\mathrm{d}y)\ln|x-y| - \frac{\sqrt{\alpha}}{2}\int \rho_{\mathbf{S}}(\mathrm{d}x)x - \frac{\alpha-1}{2}\int \rho_{\mathbf{S}}(\mathrm{d}x)\ln x \\ &-\frac{\alpha}{4}\int \rho_{\mathbf{Y}}(\mathrm{d}x)\rho_{\mathbf{Y}}(\mathrm{d}y)\ln|x-y| - \frac{\alpha}{4}\int_0^\Delta \mathrm{d}t \int \mathrm{d}x\,\rho(x,t)\left[\frac{\pi^2}{3}\rho(x,t)^2 + v(x,t)^2\right]\Bigg\}\end{aligned}}$$

Complex Burgers equation

$$\boxed{\begin{aligned} &f = v + i\pi\rho \\ &\begin{cases} \partial_t f + f\partial_x f &= 0, \\ \rho(x,t=0) &= \rho_{\mathbf{S}}(x), \\ \rho(x,t=\Delta) &= \rho_{\mathbf{Y}}(x). \end{cases}\end{aligned}}$$

The free entropy: simplifications

I Bayes-optimality

Nishimori identity

$$\mathbb{E}\left\langle \int \phi(\lambda)\rho_{\mathbf{S}}(\mathrm{d}\lambda)\right\rangle = \mathbb{E}\int \phi(\lambda)\rho^\star_{\mathbf{S}}(\mathrm{d}\lambda)$$

$$\Phi_{\mathbf{Y}} = C(\alpha) + \sup_{\rho_{\mathbf{S}}}\{\cdots\} \Rightarrow \Phi_{\mathbf{Y}} = C(\alpha) + \{\cdots\}_{\rho_{\mathbf{S}}=\rho^\star_{\mathbf{S}}}$$

II Solution to Burgers equation

Complex Burgers equation

$$f = v + i\pi\rho$$

$$\begin{cases} \partial_t f + f\partial_x f &= 0, \\ \rho(x,t=0) &= \rho_{\mathbf{S}}(x), \\ \rho(x,t=\Delta) &= \rho_{\mathbf{Y}}(x). \end{cases}$$

$\mathbf{Y}(t)/\sqrt{m} \stackrel{\mathrm{d}}{=} \mathbf{S} + \sqrt{t/m}\mathbf{Z}$ Dyson Brownian motion

$$\begin{cases} \rho(x,t) &= \frac{1}{\pi}\lim_{\epsilon\downarrow 0}\left\{\mathrm{Im}[g_{\mathbf{Y}(t)}(x+i\epsilon)]\right\}, \\ v(x,t) &= \lim_{\epsilon\downarrow 0}\left\{-\mathrm{Re}[g_{\mathbf{Y}(t)}(x+i\epsilon)]\right\}. \end{cases}$$

- We can evaluate numerically the free entropy $\Phi_{\mathbf{Y}}$ and the asymptotic denoising MMSE.
- The order parameter of the problem seems to be the probability measure $\rho_{\mathbf{S}}$!

- Coherent with our PGY approach and recent replica results of [Barbier&Macris '21].
- Very different from the low-rank case: order parameter is the $k \times k$ "overlap" matrix.

Optimal estimator

$$\mathbf{Y}/\sqrt{m} = \mathbf{S}^\star + \sqrt{\Delta/m}\mathbf{Z} = \sum_{\rho=1}^m y_\rho \mathbf{v}_\rho \mathbf{v}_\rho^\intercal$$

Bayes-optimal estimator

$$\boxed{\langle S_{\mu\nu}\rangle = \frac{Y_{\mu\nu}}{\sqrt{m}} - \Delta n \frac{\partial \Phi_{\mathbf{Y},n}}{\partial (Y_{\mu\nu}/\sqrt{m})}}$$

$\Phi_{\mathbf{Y},n}$ only depends on the spectrum of $\mathbf{Y}$

$$\langle S_{\mu\nu}\rangle = \sum_{\rho=1}^m \left[y_\rho - 2\Delta n \frac{\partial \Phi_{\mathbf{Y},n}}{\partial y_\rho}\right] v_\mu^\rho v_\nu^\rho.$$

- $\langle \mathbf{S}\rangle$ and $\mathbf{Y}$ share the same eigenvectors !
- Reduces to an optimization over the eigenvalues $\{\hat{\xi}_\mu\}$.
- Optimal RIE for denoising derived in [Bun&al '16].

$$\boxed{\langle \mathbf{S}\rangle = \hat{\mathbf{S}}_{\mathrm{RIE}} = \sum_{\mu=1}^m \hat{\xi}_\mu \mathbf{v}_\mu \mathbf{v}_\mu^\intercal}$$

Rotationally-invariant estimator (RIE)

$$\boxed{\hat{\xi}_\mu = y_\mu - 2\Delta v_{\mathbf{Y}}(y_\mu, \Delta)}$$

$$\begin{cases} v_{\mathbf{Y}}(x,\Delta) \equiv -\lim_{\epsilon\downarrow 0} \mathrm{Re}[g_{\mathbf{Y}}(x+i\epsilon)] \\ g_{\mathbf{Y}}(z) \equiv \lim_{m\to\infty} (1/m)\mathrm{Tr}[(\mathbf{Y}/\sqrt{m} - z)^{-1}] \end{cases}$$

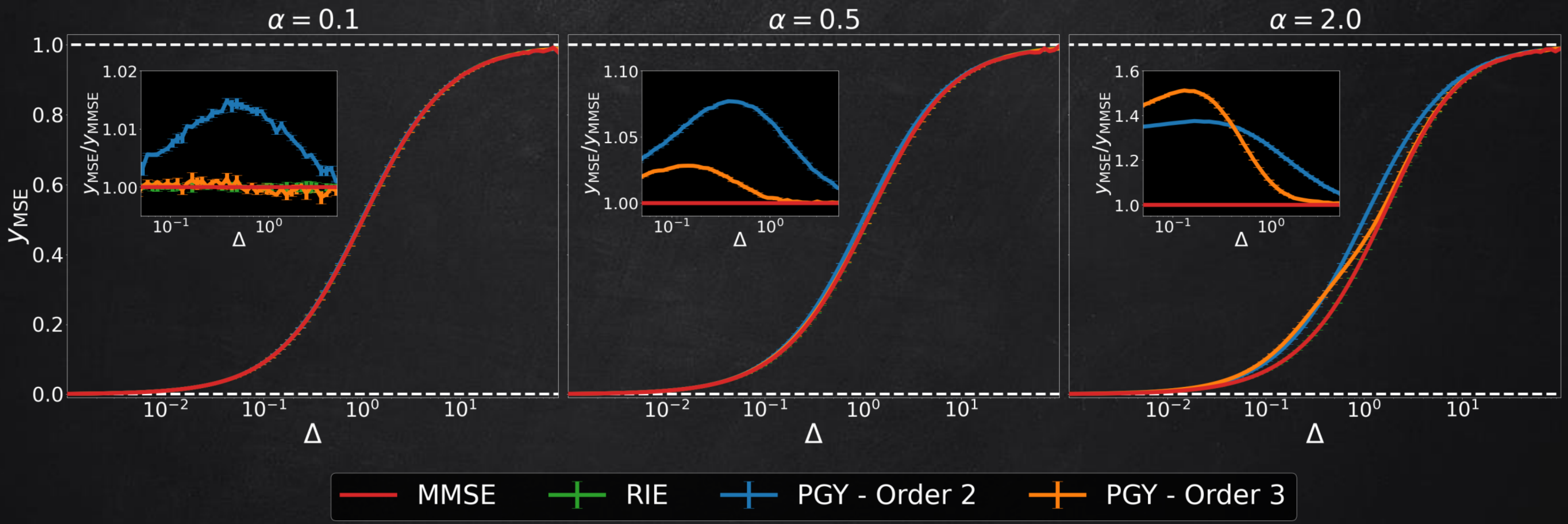
Small-α expansion: $\hat{\xi}_\mu = \frac{1}{\Delta+1} y_\mu - \frac{\Delta}{(\Delta+1)^2}\left[1 - \frac{y_\mu^2}{\Delta+1}\right]\sqrt{\alpha} + \mathcal{O}(\alpha)$ We find back the "PGY Order 3"

"Miracle": The B-O estimator only depends on the observations $\mathbf{Y}$, not on $\mathbf{S}^\star$! Numerical evaluation is easy!

Numerical recovery (1)

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

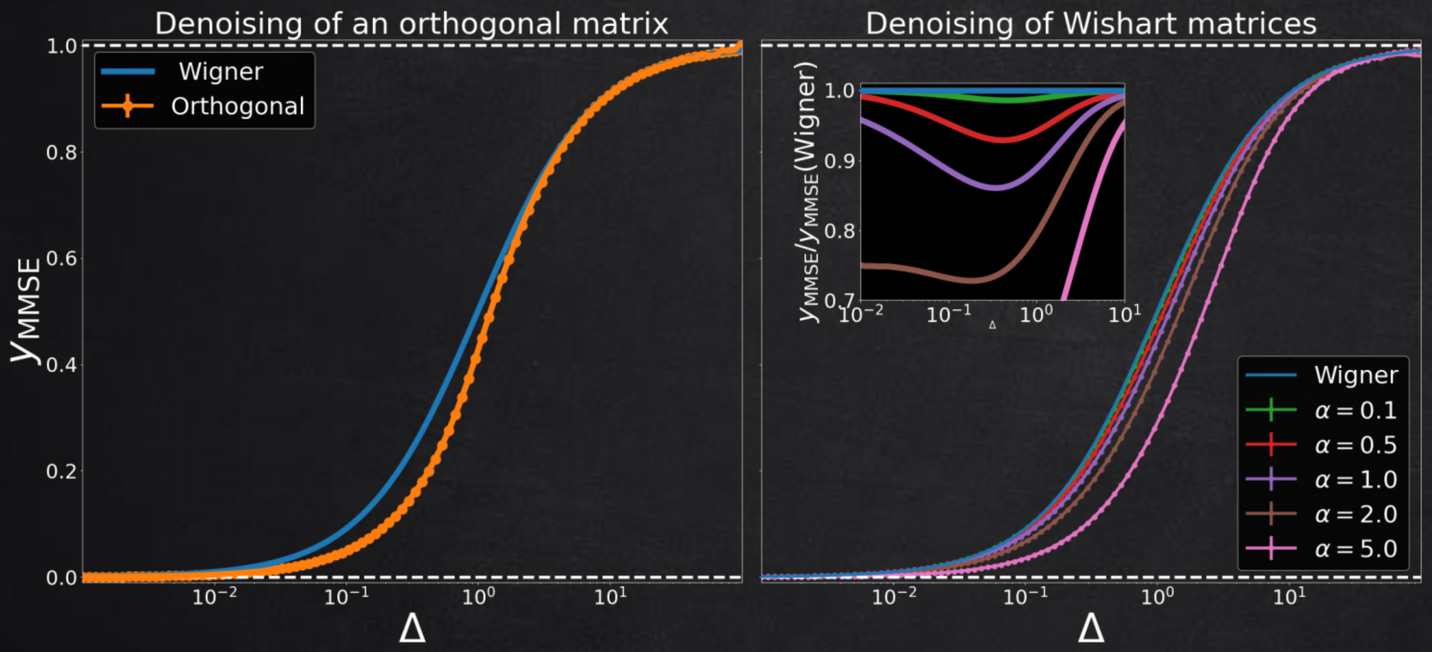
MSE of the denoising of a Wishart matrix



- Very good agreement of RIE on finite-size instances and the analytical MMSE prediction.
- "PGY order 3" significantly improves over order 2, in the overcomplete regime $\alpha \ll 1$.

Numerical recovery (2)

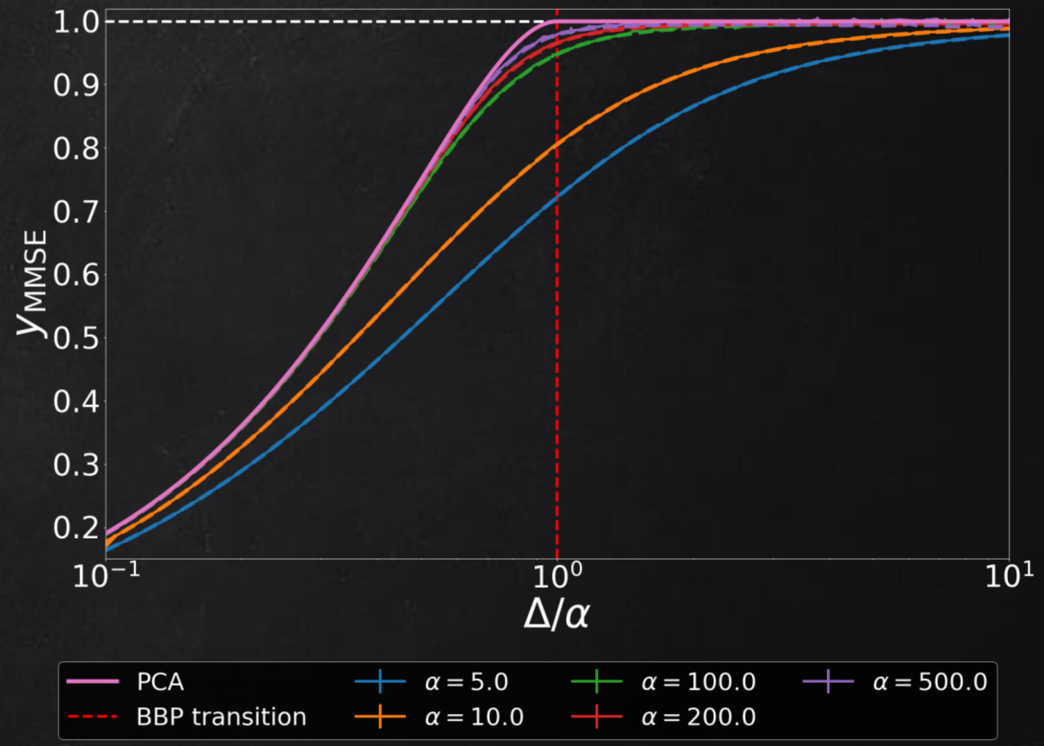
Optimal denoising for different types of $\mathbf{S}^\star$



Improvement over Wigner denoising, which increases as the matrix becomes more and more "structured".

The limit $\alpha \to \infty$

"Undercomplete" (low-rank) regime



We approach the classical 'BBP' transition

Concluding remarks

Some (of the many) open directions

- Use order-3 PGY equations as a small-α algorithm for matrix factorization ?
- PGY at all orders, and resummation of the complete series ? cf [A.M.&al '19] for low-rank models
- Denoising extensive-rank matrices with non-Gaussian noise $Y_{\mu\nu} \sim P_{\rm out}(\cdot|\sqrt{m}S^\star_{\mu\nu})$.
- Transition between low-rank and extensive-rank regimes:

$$\int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O}\, e^{\frac{\sqrt{m}}{2}\sum_{i=1}^k \theta_i (\mathbf{OYO}^\intercal)_i} \simeq \exp\left\{\frac{m}{2}\sum_{i=1}^k \int_0^{\theta_i} \mathcal{R}_{\mathbf{Y}}(-u)\mathrm{d}u\right\} \quad \Longleftrightarrow \quad \int_{\mathcal{O}(m)} \mathcal{D}\mathbf{O}\, e^{\frac{\sqrt{m}}{2\Delta}\mathrm{Tr}[\mathbf{YOLO}^\intercal]} \simeq \exp\left\{\frac{m^2}{2} I_\Delta[\rho_{\mathbf{Y}}, \rho_{\mathbf{S}}]\right\}$$

$k = \Theta(1)$? $k = \Theta(m)$

THANK YOU !