Matrix Dyson Equation for Correlated Linearizations

Hugo Latourelle-Vigeant

M.Sc. student, McGill University, Department of Mathematics and Statistics

The many facets of random matrix theory, 2023 CMS Winter Meeting

McGill

Outline

1. Background
2. Framework
3. Application: Random features

Background

Given $H \in \mathbb{R}^{n\times n}$ a self-adjoint matrix, consider the *matrix Dyson equation (MDE)*

$$(\mathbb{E}H - \underbrace{\mathbb{E}_{\widetilde{H}}[(\widetilde{H} - \mathbb{E}H)M(\widetilde{H} - \mathbb{E}H)]}_{:=\mathcal{S}(M) \text{ (superoperator)}} - zI_n)M = I_n$$

where $\Im[z] > 0$ and $\Im[M] \succ 0$.

Properties of the MDE

- There exists a unique analytic solution M to the MDE [HFS07]

Properties of the MDE

- There exists a unique analytic solution M to the MDE [HFS07]
- $\|M(z)\| \leq 1/\Im[z]$, Stieltjes transform representation, etc.

Properties of the MDE

- There exists a unique analytic solution M to the MDE [HFS07]
- $\|M(z)\| \leq 1/\Im[z]$, Stieltjes transform representation, etc.
- Under some assumptions, if the entries of H are "weakly correlated", $(H - zI_n)^{-1} \approx M(z)$ in the sense of isotropic and averaged local laws

Properties of the MDE

- There exists a unique analytic solution M to the MDE [HFS07]
- $\|M(z)\| \leq 1/\Im[z]$, Stieltjes transform representation, etc.
- Under some assumptions, if the entries of H are "weakly correlated", $(H - zI_n)^{-1} \approx M(z)$ in the sense of isotropic and averaged local laws
- By "weakly correlated", I mean a generalization of Wigner matrices
  - If H is Wigner, then $\mathcal{S}(M) \approx \frac{c\,\mathbf{tr}(M)}{n} I$
  - If H is Wishart, then $\mathcal{S}(M) \approx \frac{c\,\mathbf{tr}(M)}{n} I$

How can we use the matrix Dyson equation framework to study, for instance, Wishart matrices?

Linearization trick

Linearization trick (Belinschi, Mai, and Speicher '13)

Let p be a self-adjoint n by n polynomial expression in $\mathbb{C}\langle X_1, \ldots, X_k\rangle$. Then, there exists a linearization $L \in \mathbb{C}^{(n+d)\times(n+d)}$ such that

1. L is linear in $X_1, X_2, \ldots, X_k$
2. $(L - z\Lambda)^{-1}_{1\leq i,j\leq n} = (p - zI_n)^{-1}$ where $\Lambda = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}$
3. $L = \begin{bmatrix} A & B^* \\ B & Q \end{bmatrix}$ with Q invertible.

Linearizations are also called (affine) pencils or realizations.

Examples

- (Gram matrix)

$$\begin{bmatrix} -zI & X \\ X^* & -I \end{bmatrix}^{-1}_{1,1} = (XX^* - zI)^{-1}$$

Examples

- (Gram matrix)

$$\begin{bmatrix} -zI & X \\ X^* & -I \end{bmatrix}^{-1}_{1,1} = (XX^* - zI)^{-1}$$

- (Sample covariance matrix)

$$\begin{bmatrix} -zI & 0 & 0 & X \\ 0 & 0 & Y & -I \\ 0 & Y^* & -I & 0 \\ X^* & -I & 0 & 0 \end{bmatrix}^{-1}_{1,1} = (XYY^*X^* - zI)^{-1}$$

Examples

- (Gram matrix)

$$\begin{bmatrix} -zI & X \\ X^* & -I \end{bmatrix}^{-1}_{1,1} = (XX^* - zI)^{-1}$$

- (Sample covariance matrix)

$$\begin{bmatrix} -zI & 0 & 0 & X \\ 0 & 0 & Y & -I \\ 0 & Y^* & -I & 0 \\ X^* & -I & 0 & 0 \end{bmatrix}^{-1}_{1,1} = (XYY^*X^* - zI)^{-1}$$

- (Anticommutator)

$$\begin{bmatrix} -zI & X & Y \\ X^* & 0 & -I \\ Y^* & -I & 0 \end{bmatrix}^{-1}_{1,1} = (XY^* + YX^* - zI)^{-1}$$

Linearization algorithmically

$$
Q^{K^{-1}} = \begin{pmatrix}
I_m & \frac{\sqrt{\eta-\zeta}\Theta^\top}{\gamma\sqrt{n_1}} & \frac{\sqrt{\rho}X^\top}{\gamma\sqrt{n_0}} & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{\Theta\sqrt{\eta-\zeta}}{\sqrt{n_1}} & I_{n_1} & 0 & 0 & -\frac{\sqrt{\rho}W}{\sqrt{n_1}} & 0 & 0 & 0 & 0 \\
0 & 0 & I_{n_0} & -\Sigma^{1/2} & 0 & 0 & 0 & 0 & 0 \\
0 & -\frac{W^\top}{\sqrt{n_1}} & 0 & I_{n_0} & 0 & 0 & \frac{\Sigma_\beta}{\sqrt{\rho}} & 0 & 0 \\
0 & 0 & 0 & 0 & I_{n_0} & -\Sigma^{1/2} & 0 & 0 & 0 \\
-\frac{X}{\sqrt{n_0}} & 0 & 0 & 0 & 0 & I_{n_0} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{n_0} & -\Sigma^{1/2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{n_0} & -\frac{X}{\sqrt{n_0}} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_m
\end{pmatrix}. \tag{S171}
$$

Figure 1: Linearization obtained algorithmically. Source: “Anisotropic random feature regression in high dimensions” by Mel and Pennington

Pseudo-resolvents & MDE

The linearization trick leads to the study of *pseudo-resolvents* $(L - z\Lambda)^{-1}$.

The matrix Dyson equation framework has been adapted to analyze pseudo-resolvents:

- On a global scale (Anderson '13)
- On a local scale (Erdős, Krüger, and Nemish '18)

Those work used free probability, and apply to linearizations with generalized Wigner and/or non-symmetric random matrices.

Our goal

1. Extend the matrix Dyson equation framework to derive anisotropic global laws for pseudo-resolvents of linearizations with arbitrary correlation structure
2. Present motivating example from machine learning

Framework

Settings

We are given

- A linearization $L = \begin{bmatrix} A & B^T \\ B & Q \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}$,
  - $A \in \mathbb{R}^{n \times n}$ self-adjoint, $B \in \mathbb{R}^{d \times n}$
  - $Q \in \mathbb{R}^{d \times d}$ invertible, self-adjoint and deterministic
- $\Lambda = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix}$
- A spectral parameter $z \in \mathbb{C}$ with $\Im[z] > 0$

We want to find a deterministic equivalent $(L - z\Lambda)^{-1}$

Matrix Dyson equation for (correlated) linearizations

Consider the MDE

$$(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)M = I_\ell$$

with

- superoperator $\mathcal{S}(M) := \mathbb{E}[(L - \mathbb{E}L)M(L - \mathbb{E}L)] - \tilde{\mathcal{S}}(M)$
- $\tilde{\mathcal{S}}(M) = \begin{bmatrix} 0 & \mathbb{E}[(B - \mathbb{E}B)^T M_{2,1} (B - \mathbb{E}B)^T] \\ \mathbb{E}[(B - \mathbb{E}B) M_{1,2} (B - \mathbb{E}B)] & 0 \end{bmatrix}$

Existence and uniqueness

Because the spectral parameter does not span the entire diagonal, existence of a solution to the MDE is not trivial.

Existence and uniqueness

Because the spectral parameter does not span the entire diagonal, existence of a solution to the MDE is not trivial.

Define the *admissible set*

$$\mathscr{A} = \{W \in \mathbb{C}^{\ell\times\ell} \,:\, \Im[(W)_{i,j=1}^{n}] \succ 0,\, \Im[W] \succeq 0\}$$

Theorem (L.V., Paquette '23)

There exists a unique analytic $M : \mathbb{H} \mapsto \mathscr{A}$ that solves the MDE.

This $M(z)$ is the candidate deterministic equivalent for $(L - z\Lambda)^{-1}$.

Regularized matrix Dyson equation

- **Problem**: It is difficult to work directly with $(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)M = I_\ell$

Regularized matrix Dyson equation

- **Problem**: It is difficult to work directly with $(\mathbb{E}L - \mathcal{S}(M) - z\Lambda)M = I_\ell$
- **Solution**: For every $\tau > 0$, define the *regularized matrix Dyson equation* (RMDE)

$$(\mathbb{E}L - \mathcal{S}(M^{(\tau)}) - z\Lambda - i\tau I_\ell)M^{(\tau)} = I_\ell$$

  and an admissible set $\mathcal{A}_+ = \{W \in \mathbb{C}^{\ell\times\ell} : \Im[W] \succ 0\}$.
  - Unique analytic $M^{(\tau)} : \mathbb{H} \mapsto \mathcal{A}_+$ solution to the RMDE
  - $W \mapsto (\mathbb{E}L - \mathcal{S}(W) - z\Lambda - i\tau I_\ell)^{-1}$ contraction in CRF-pseudometric
  - We define $\lim_{\tau\to 0} M^{(\tau)}(z) = M(z)$

Regularized pseudo-resolvent

The expected regularized pseudo-resolvent almost solves the RMDE up to an additive perturbation $D^{(\tau)}$:

$$(\mathbb{E}L - \mathcal{S}(\mathbb{E}(L - z\Lambda - i\tau)^{-1}) - z\Lambda - i\tau I_\ell)\mathbb{E}(L - z\Lambda - i\tau)^{-1} = I_\ell + D^{(\tau)}$$

with

$$D^{(\tau)} = \mathbb{E}\left[\left(\mathbb{E}L - L - \mathcal{S}(\mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1})\right)(L - z\Lambda - i\tau I_\ell)^{-1}\right].$$

Stability

Theorem (L.-V., Paquette '23)

If

- $\|M^{(\tau)}(z) - M(z)\| \xrightarrow{\tau \to 0} 0$ *uniformly in* ℓ
- $\|\mathcal{S}\|$, $\|\mathbb{E}L\|$ *and* $\mathbb{E}\|(L - z\Lambda)^{-1}\|^2$ *are bounded.*
- $\|D^{(\tau)}\| \xrightarrow{\ell \to \infty} 0$ *for every* $\tau > 0$

then $\|M(z) - \mathbb{E}(L - z\Lambda)^{-1}\| \xrightarrow{\ell \to \infty} 0$ *for every* $z \in \mathbb{H}$.

$$(L - z\Lambda)^{-1} \approx \mathbb{E}(L - z\Lambda)^{-1} \approx \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}$$

$$M(z) \approx M^{(\tau)}(z) \approx \mathbb{E}(L - z\Lambda - i\tau I_\ell)^{-1}$$

Assumption: $\|M^{(\tau)}(z) - M(z)\| \xrightarrow{\tau \to 0} 0$ uniformly in ℓ

- We need $\|M^{(\tau)}(z) - M(z)\| \xrightarrow{\tau \to 0} 0$ uniformly in ℓ
- Ensures stability of the MDE
- When L has Wigner blocks, we can use free semicircular variables to construct a dimension independent representation of M and $M^{(\tau)}$ [And13; EKN18; FKN23]

Assumption: $\|D^{(\tau)}\| \xrightarrow{\ell\to\infty} 0$ for every $\tau > 0$

Theorem (L.-V., Paquette 23')

If $L \equiv L(g) = \mathcal{C}(g) + \mathbb{E}L$ for some $g \sim \mathcal{N}(0, I_\gamma)$, then

$$\|D^{(\tau)}\| \leq c\tau^{-1}\sqrt{\ell}\lambda + \tau^{-2}\|\tilde{\mathcal{S}}\| + \|\Delta(L,\tau)\|$$

with

- $g \mapsto \mathcal{S}\left((L(g) - z\Lambda - i\tau I_\ell)^{-1}\right)$ *is λ-Lipschitz with respect to the operator norm*
- $\tilde{\mathcal{S}}$ *is the part that we removed from $\mathcal{S}$*
- $\|\Delta(L,\tau)\|$ *relates to how close L is to satisfying a matrix Stein's lemma*

Application: Random features

Setup

- Dataset $\{(x_j, y_j)\}_{j=1}^{n_{train}}$ with $x_j \in \mathbb{R}^{n_0}$ and $y_j \in \mathbb{R}$
- Want to learn relation between x_j and y_j using

$$\min_{w \in \mathbb{R}^d} \|y - Aw\|^2 + \delta \|w\|^2$$

  - $A = n^{-\frac{1}{2}} \sigma(XW)$
  - $W \in \mathbb{R}^{n_0 \times d}$ is a matrix of i.i.d. Gaussians
  - σ Lipschitz functions
  - ridge parameter $\delta > 0$
  - $\mathbb{E}A = 0$
- Explicit solution $w = A^T(AA^T + \delta I_{n_{train}})^{-1} y$

Why study random features?

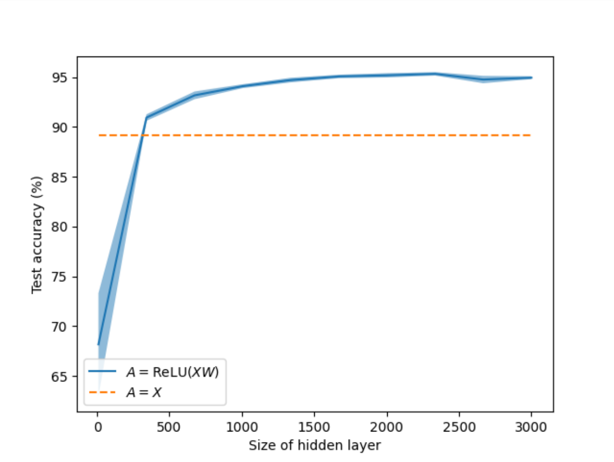

Figure 2: Average and standard deviation over 10 runs of even/odd classification of MNIST using a random feature model. $n_{train} = 6000$, $n_{test} = 10000$ and $\delta = 0.01$.

Random features as a null model

Random features serves as a toy model for neural networks

- Double/multiple descents (Mei and Montanari '19)
- Implicit regularization (Jacot et al. '20)
- Universality (Hu and Lu '20)

Test error

Given other dataset $\{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{n_{test}}$ with $\hat{x}_j \in \mathbb{R}^{n_0}$ and $\hat{y}_j \in \mathbb{R}$, the test error is

$$E_{test} := \|\widehat{y} - \widehat{A}w\|^2 = \|\widehat{y} - \widehat{A}A^T(AA^T + \delta I_{n_{train}})^{-1}y\|^2$$

with $\widehat{A} = n^{-\frac{1}{2}}\sigma(\widehat{X}W) \in \mathbb{R}^{n_{test} \times d}$.

Linearization

$$L = \begin{bmatrix} \delta I_{n_{\mathrm{train}}} & A & 0_{n_{\mathrm{train}} \times n_{\mathrm{test}}} & 0_{n_{\mathrm{train}} \times n_{\mathrm{test}}} \\ A^T & -I_{d \times d} & 0_{d \times n_{\mathrm{test}}} & \widehat{A}^T \\ 0_{n_{\mathrm{test}} \times n_{\mathrm{train}}} & 0_{n_{\mathrm{test}} \times d} & 0_{n_{\mathrm{test}} \times n_{\mathrm{test}}} & -I_{n_{\mathrm{test}}} \\ 0_{n_{\mathrm{test}} \times n_{\mathrm{train}}} & \widehat{A} & -I_{n_{\mathrm{test}}} & 0_{n_{\mathrm{test}} \times n_{\mathrm{test}}} \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}.$$

Taking $\Lambda := \mathrm{BlockDiag}\{I_{n_{\mathrm{train}}+d}, 0_{2n_{\mathrm{test}} \times 2n_{\mathrm{test}}}\}$, we form the pseudo-resolvent $(L - z\Lambda)^{-1}$ and we get

$$(L - z\Lambda)^{-1}_{3,1} = (1+z)^{-1} \widehat{A} A^T \left((1+z)^{-1} A A^T + (\delta - z) I_{n_{\mathrm{train}}} \right)^{-1}.$$

Main result

Theorem (L.-V., Paquette '23)

Assume that $n_{\text{train}}, d, n_{\text{test}}, n_0 \propto n$ *and* $\mathbb{E}[\|A\|^4], \mathbb{E}[\|\widehat{A}\|^4]$ *are bounded. Let* α *be the unique non-positive real number satisfying*

$$\alpha = -\left(1 + \operatorname{tr}\left(K_{AA^T}(\delta I_{n_{train}} - d\alpha K_{AA^T})^{-1}\right)\right)^{-1} \in \mathbb{R}_{\leq 0}$$

and denote $M = (\delta I_{n_{\text{train}}} - d\alpha K_{AA^T})^{-1}$ *as well as*

$$\beta = \frac{\alpha^2 \operatorname{tr}\left(K_{\widehat{A}\widehat{A}^T} + d\alpha K_{\widehat{A}A^T} M (I_{n_{\text{train}}} + \delta M) K_{A\widehat{A}^T}\right)}{1 - \|\sqrt{d}\alpha K_{AA^T}^{\frac{1}{2}} M K_{AA^T}^{\frac{1}{2}}\|_F^2} \in \mathbb{R}_{\geq 0}.$$

Then, $E_{\text{test}} \xrightarrow[n\to\infty]{a.s.} d\beta \|K_{AA^T}^{\frac{1}{2}} My\|^2 + \|d\alpha K_{\widehat{A}A^T} My + \widetilde{y}\|^2.$

Here, K_{AA^T}, $K_{\widehat{A}A^T}$ and $K_{\widehat{A}\widehat{A}^T}$ are covariance matrices.

Gaussian equivalence theorem

As a consequence, we may replace a random features model by an equivalent surrogate Gaussian matrix with matching covariance.

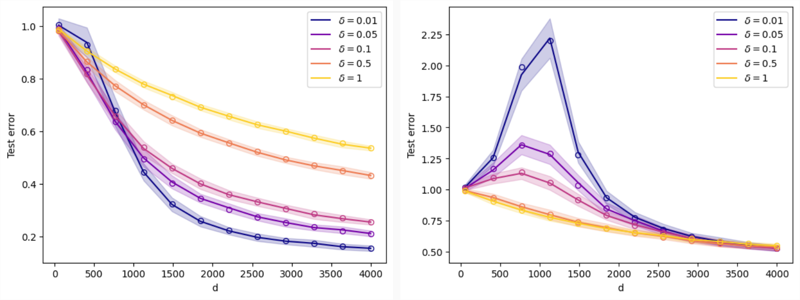

Figure 3: E_{test} vs the deterministic approximation for various odd activation functions with different size of hidden layers d and ridge parameter δ. Left: Error function activation ($\sigma(x) = \mathrm{erf}(x)$); Right: Sign activation ($\sigma(x) = \mathrm{sign}(x)$).

References

[And13] Greg W. Anderson. **“Convergence of the largest singular value of a polynomial in independent Wigner matrices”.** In: *The Annals of Probability* 41.3B (May 2013). DOI: 10.1214/11-aop739.

[BMS17] Serban T. Belinschi, Tobias Mai, and Roland Speicher. **“Analytic subordination theory of operator-valued free additive convolution and the solution of a general random matrix problem”.** In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 2013.732 (2017), pp. 21–53. DOI: doi:10.1515/crelle-2014-0138.

[EKN18] László Erdős, Torben Krüger, and Yuriy Nemish. **“Local laws for polynomials of Wigner matrices”.** In: *Journal of Functional Analysis* 278.12 (2018), p. 108507. ISSN: 0022-1236. DOI: 10.1016/j.jfa.2020.108507.

[FKN23] Jacob Fronk, Torben Krüger, and Yuriy Nemish. ***Norm Convergence Rate for Multivariate Quadratic Polynomials of Wigner Matrices.*** 2023. arXiv: 2308.16778 [math.PR].

[HFS07] J. William Helton, Reza Rashidi Far, and Roland Speicher. ***Operator-valued semicircular elements: Solving a quadratic matrix equation with positivity constraints.*** 2007. arXiv: math/0703510 [math.OA].

[HL22] Hong Hu and Yue M. Lu. **"Universality Laws for High-Dimensional Learning with Random Features".** In: *IEEE Transactions on Information Theory, in press* (2022).

[Jac+20] Arthur Jacot et al. **“Implicit Regularization of Random Feature Models”.** In: *Proceedings of the 37th International Conference on Machine Learning*. ICML’20. JMLR.org, 2020.

[MM22] Song Mei and Andrea Montanari. **“The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve”.** In: *Communications on Pure and Applied Mathematics* 75.4 (2022), pp. 667–766. DOI: `https://doi.org/10.1002/cpa.22008`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.22008`.

[MP21] Gabriel Mel and Jeffrey Pennington. **“Anisotropic random feature regression in high dimensions”.** In:

International Conference on Learning Representations. 2021.

Addendum

Stability operator

The stability operator is defined as

$$\mathcal{L} : W \in \mathbb{C}^{\ell\times\ell} \mapsto W - M\mathcal{S}(W)M.$$

It is related to our assumption $\|M(z) - M^{(\tau)}(z)\| \xrightarrow{\tau\to 0} 0$ uniformly in ℓ because

$$\mathcal{L}(\partial_{i\tau} M(z)) = (M(z))^2.$$

Expansion of test error

$$
\begin{aligned}
E_{test} &:= \|\widetilde{y} - \widetilde{A}\beta\|^2 \\
&= -2\widetilde{y}^T \underbrace{\widetilde{A}A^T \left(AA^T + \delta I_{n_{train}}\right)^{-1} y}_{(1)} \\
&+ y^T \underbrace{(AA^T + \delta I_{n_{train}})^{-1} A\widetilde{A}^T \widetilde{A}A^T (AA^T + \delta I_{n_{train}})^{-1} y}_{(2)} \\
&+ \|\widetilde{y}\|^2
\end{aligned}
$$

Superoperator

The linearization presented for the motivating example huge, but it has a simple correlation structure:

$$\mathcal{S}^{(1)}(M) = \begin{bmatrix} \mathrm{tr}(M_{2,2})XX^T & 0 & 0 & \mathrm{tr}(M_{2,2})X\widetilde{X}^T \\ 0 & \rho(M)I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \mathrm{tr}(M_{2,2})\widetilde{X}X^T & 0 & 0 & \mathrm{tr}(M_{2,2})\widetilde{X}\widetilde{X}^T \end{bmatrix}$$

where $\rho(M) := \mathrm{tr}(XX^T M_{1,1} + X\widetilde{X}^T M_{4,1} + \widetilde{X}X^T M_{1,4} + \widetilde{X}\widetilde{X}^T M_{4,4})$

Applying our framework

1. There is a unique solution M to the associated MDE

$$M(z) = \begin{bmatrix} ((\delta - z)I_{n_{\mathrm{train}}} - \operatorname{tr}(M_{2,2})K_{AA^T})^{-1} & 0 & -\operatorname{tr}(M_{2,2})M_{1,1}K_{A\widehat{A}^T} & 0 \\ 0 & -(1 + z + \operatorname{tr}(K_{AA^T}M_{1,1}))^{-1}I_d & 0 & 0 \\ -\operatorname{tr}(M_{2,2})K_{\widehat{A}A^T}M_{1,1} & 0 & (\operatorname{tr}(M_{2,2}))^2 K_{\widehat{A}A^T}M_{1,1}K_{A\widehat{A}^T} + \operatorname{tr}(M_{2,2})K_{\widehat{A}\widehat{A}^T} & -I_{n_{\mathrm{test}}} \\ 0 & 0 & -I_{n_{\mathrm{test}}} & 0 \end{bmatrix}.$$

2. We can control $\|M^{(\tau)}(z) - M(z)\|$ using the structure of M
3. To show $\|D^{(\tau)}\| \to 0$, we use

$$\|D^{(\tau)}\| \leq c\tau^{-1}\sqrt{\ell}\underbrace{\lambda}_{O(\ell^{-1})} + \tau^{-2}\underbrace{\|\tilde{\mathcal{S}}\|}_{O(\ell^{-1/2})} + \underbrace{\|\Delta(L,\tau)\|}_{LOO}$$

It only remains to take $z \to 0$...

First deterministic equivalent

Lemma

Under some boundedness assumptions,

$$\operatorname{tr}\left(U(L^{-1} - M(0))\right) \xrightarrow[n\to\infty]{a.s.} 0$$

for every $U \in \mathbb{C}^{\ell\times\ell}$ *with* $\|U\|_* \leq 1$.

Second deterministic equivalent

- Now, we want to find a deterministic equivalent for $(AA^T + \delta I_{n_{train}})^{-1} A\widetilde{A}^T \widetilde{A} A^T (AA^T + \delta I_{n_{train}})^{-1}$
- This is the "square" of the previous expression
- We can use contour integral trick along with stability of $M^{(\tau)}$
- We can extract more information about $M^{(\tau)}$, which already used to find the first deterministic equivalent, and use a contour integral trick to find the second deterministic equivalent

Numerical simulations

We only have to compute the scalar a:

Numerically solving for a

Let $a_0 \in \mathbb{R}_{<0}$ and consider the iterates

$$a_{k+1} = -\left(1 + \operatorname{tr}\left(K_{AA^T}(\delta I_{n_{\mathrm{train}}} - a_k d K_{AA^T})^{-1}\right)\right)^{-1}.$$

Then, $a = \lim_{k\to\infty} a_k$.