Introduction to the Matrix Dyson Equation

Hugo Latourelle-Vigeant

March 23, 2026

This document contains notes from an introductory lecture on the Matrix Dyson Equation, delivered at ETH in March 2024.

1 Motivation

The contemporary landscape of machine learning is marked by the vast scale of models. Ubiquitous machine learning models are trained on copious amounts of high-dimensional samples, frequently containing as many, if not more, parameters. For instance, the GPT-3 model, a precursor to the widely-used conversational agent ChatGPT, stands out with an impressive 175 billion parameters and training on 374 billion data points [Bro+20]. Additionally, the image generation systems DALL-E and DALL-E 2, capable of producing photo-realistic images from text prompts, leverages a staggering 12 billion and 3.5 billion parameters respectively, trained on 250 million and 650 million data points respectively [Ram+21; Ram+22].

This escalation in model complexity raises critical questions about the theoretical underpinnings of such expansive models. This inquiry is driven by the so-called "curse of dimensionality," a term coined by Bellman in his exposition on dynamic programming [Bel10]. The concept is loosely used to convey the idea that low-dimensional intuition, as well as computational or theoretical approaches, may break down entirely in high-dimensional scenarios [CL22]. Manifestations of this curse are observable in various aspects in the context of machine learning. As a first example, traditional performance metrics for optimization techniques often rely on worst-case complexity. Yet, in real-world, high-dimensional scenarios, the likelihood of encountering a data point associated with the worst-case running time of a given algorithm may be exceedingly low. Therefore, while worst-case complexity offers assurances regarding running time, it may not accurately reflect the algorithm's expected performance in high dimensions. A second example is found in the analysis of artificial neural networks. When overparametrized, meaning the number of trainable parameters exceeds the number of data points, these models possess extraordinary capacity. In fact, they are capable of fitting given data perfectly, even when the labels are pure noise [Zha+21]. In this scenario, the fact that they still exhibit good generalization performance contravenes conventional statistical knowledge. Furthermore, the presence of non-linear activation functions adds complexity, making them challenging to analyze analytically.

Fortunately, the scale of contemporary models not only brings about the curse of dimensionality but also offers a blessing. It turns out that, in many problems of interests, we are interested in understanding the asymptotic behavior of scalar observations of large random systems. Conceptualizing the problem as random, we may leverage the fact that scalar observations of large random systems often follow a law of large numbers effect and concentrate around a deterministic quantity. If the random distribution is chosen such that this quantity accurately characterizes observed behavior in practice, one can argue that the chosen model is satisfactory. To describe this deterministic limit, we can turn to tools from random matrix theory.

1.1 Random Matrix Theory

How do we study a sequence of random matrices that are increasing in size? First, instead of studying the random matrix directly, we will study the *resolvent*.

Definition 1. Given a Hermitian matrix $H \in \mathbb{C}^{n\times n}$, its resolvent is defined by $R(z;H) = (H - zI_n)^{-1}$.

The resolvent is an essential tool in random matrix theory because it exhibits valuable properties that encode virtually all the information about the spectrum of the underlying matrix. It is useful to think of the resolvent as an analogue to the characteristic function in probability theory.

Proposition 1. *Let $H \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Then, the resolvent $\mathcal{R}(z;H)$ is a holomorphic function on $\mathbb{C} \setminus \sigma(H)$ and $\|\mathcal{R}(z;H)\| \leq \operatorname{dist}(z,\sigma(H))^{-1} \leq (\Im[z])^{-1}$ for every $z \in \mathbb{C} \setminus \sigma(H)$. Furthermore, if $H = U\Lambda U^T$ for some orthonormal matrix U and diagonal matrix Λ and Γ is a positively oriented simple closed curve with interior Γ°, then*

$$U \operatorname{diag}\{f(\lambda_j(H))\mathbb{I}_{\lambda_j(H)\in\Gamma^\circ}\}U^T = \frac{1}{2\pi i}\oint_\Gamma f(z)\mathcal{R}(z;H)\mathrm{d}z$$

for every complex function f analytic on a region containing Γ and its inside.

Since we are interested in scalar observations of large random matrices, we define the notion of (asymptotic) *deterministic equivalence*.

Definition 2. Given a potential random matrix $A \in \mathbb{C}^{n\times n}$, we say that $B \in \mathbb{C}^{n\times n}$ is a deterministic equivalent of A if B is a deterministic matrix and $\operatorname{tr} U(A-B) \to 0$ almost surely as $n \to \infty$ for any sequence of unit matrices $U \in \mathbb{R}^{n\times n}$ with $\|U\|_* \leq 1$.

In particular, combining the notion of deterministic equivalence with the notion of resolvent, we will be interested in finding deterministic equivalent for resolvent of random matrices.

It is important to note that, traditionally, random matrix theory has been concerned with the characterizing the distribution of the eigenvalues. In particular, many results characterize the empirical spectral distribution of random matrices. This is what we would call an

isotropic/averaged law. The notion of deterministic equivalence that we presented, however, is more general. Indeed, it allows us to recover information about the alignment of the eigenvectors with some deterministic vector for instance. For this reason, the type of results which have this type of convergence are called *anisotropic laws*.

2 Matrix Dyson Equation

There exists multiple methods to derive deterministic equivalents for resolvents of random matrices. For instance, one can use the moment methods, which involves more combinatorics computations. There is also the leave-one-out method, which is simple and powerful but requires a good "guess" regarding the deterministic equivalent. In this lecture, we will focus on the matrix Dyson equation.

Instead of just stating the matrix Dyson equation without any context, we provide a heuristic derivation for GOE matrix. Let $G \in \mathbb{R}^{n\times n}$ be a GOE matrix. That is, G is a symmetric matrix with $G_{i,j} \sim \mathcal{N}(0,1)$ for $i \neq j$ and $G_{i,i} \sim \mathcal{N}(0,2)$. In particular, $G = 2^{-1/2}(Z + Z^T)$ where Z is a matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Let $H = n^{-\frac{1}{2}}G$ be a normalized GOE matrix. We write

$$\mathbb{E}[G(H - zI_n)^{-1}]_{j,k} = \sum_{a\neq j} \mathbb{E}[G_{j,a}(H - zI_n)^{-1}_{a,k}] + \mathbb{E}[G_{j,j}(H - zI_n)^{-1}_{j,k}].$$

Let $E_{j,k}$ is the matrix which has all zero entries except the (j,k) entry, which is 1. By Stein's lemma,

$$\mathbb{E}[G_{j,a}(H-zI_n)^{-1}_{a,k}] = \mathbb{E}\left[\frac{\partial(H - zI_n)^{-1}}{\partial G_{j,a}}\right]_{a,k} = -n^{-\frac{1}{2}}\mathbb{E}[(H-zI_n)^{-1}(E_{j,a}+E_{a,j})(H-zI_n)^{-1}]_{a,k}.$$

Hence, we have

$$\begin{aligned}\mathbb{E}[H(H - zI_n)^{-1}] &= -n^{-1}\mathbb{E}[\mathrm{tr}((H - zI_n)^{-1})(H - zI_n)^{-1}] - n^{-1}\mathbb{E}[(H - zI_n)^{-2}] \\ &\quad + 2n^{-1}\mathbb{E}[\mathrm{diag}\{(H - zI_n)^{-1}_{j,j}\}_{j=1}^n(H - zI_n)^{-1}] + n^{-\frac{1}{2}}\mathbb{E}[\mathrm{diag}\{G_{j,j}\}_{j=1}^n(H - zI_n)^{-1}]\end{aligned}$$

Using Gaussian concentration for instance, $n^{-1}\,\mathrm{tr}((H-zI_n)^{-1})$ concentrates around its mean. Hence,

$$\mathbb{E}[H(H - zI_n)^{-1}] = -\mathbb{E}[n^{-1}\,\mathrm{tr}(H - zI_n)^{-1}]\mathbb{E}[(H - zI_n)^{-1}] + o_n(1).$$

Using the structure of the GOE,

$$\mathbb{E}[n^{-1}\,\mathrm{tr}(H - zI_n)^{-1}] = \mathbb{E}[H\mathbb{E}[(H - zI_n)^{-1}]H] + o_n(1).$$

To summarize, using the tautological identity $I_n = (H-zI_n)(H-zI_n)^{-1}$ and the previous computations, we have shown that

$$-(\mathbb{E}[H\mathbb{E}[(H - zI_n)^{-1}]H] + zI_n)\mathbb{E}(H - zI_n)^{-1} = I_n + o_n(1).$$

The expected resolvent is a deterministic equivalent for the resolvent itself, and by universality we expect that this result holds for more general distributions. This is the matrix Dyson equation.

Definition 3. Given a symmetric random matrix $H \in \mathbb{R}^{n\times n}$, the matrix Dyson equation takes the form

$$(\mathbb{E}H - \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)] - zI_n)M = I_n$$

for $z \in \mathbb{H}$. The equation contains the mean matrix $\mathbb{E}H$ and encodes the covariance structure in the so-called superoperator $M \mapsto \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)]$.

We give an overview of the approach used to derive deterministic equivalents using the matrix Dyson equation.

2.1 Existence and Uniqueness

The first step is to show the existence of a unique solution to the matrix Dyson equation. To do so, we use the literature on fixed point theory. Let

$$\mathscr{M} := \{f : \mathbb{H} \mapsto \mathscr{A} \text{ analytic}\}, \quad \mathscr{A} := \{M \in \mathbb{C}^{n\times n} : \Im[M] \succ 0\}$$

be the *admissible set*. We look for solutions in the admissible set. Define the MDE map:

$$\mathcal{F} : M \in \mathscr{M} \mapsto (\mathbb{E}H - \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)] - zI_n)^{-1} \in \mathscr{M}.$$

The main property of the MDE mapping $\mathcal{F}$ is that it maps the admissible set strictly into itself. Let $\mathscr{M}_b := \{f : \{z \in \mathbb{H} : |z| \le b\} \mapsto \mathscr{A} \cap \mathbb{B}_b \text{ analytic}\}$ for every $b \in \mathbb{R}_{>0}$.

Proposition 2. *Let $z \in \mathbb{H}$, $b \in \mathbb{R}_{>0}$ and define $m_b = \|\mathbb{E}H\| + \mathbb{E}\|H-\mathbb{E}H\|^2 b + |z|$. Then, for every $M \in \mathscr{A}$ with $\|M\| \le b$,*

$$\|\mathcal{F}^{(\tau)}(M)\| \le (\Im[z])^{-1}, \qquad \Im[\mathcal{F}^{(\tau)}(M)] \succeq \Im[z] m_b^{-2} I_\ell \succ 0.$$

In particular, if $b > \Im[z]$, then $\mathcal{F}^{(\tau)}$ maps $\mathscr{M}_b$ strictly into itself.

Proof. First of all, we have

$$\Im[\mathcal{F}(M)] = -\mathcal{F}(M)\Im[\mathbb{E}H - \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)] - zI_n]\mathcal{F}(M)^* \succeq \Im[z]\mathcal{F}(M)\mathcal{F}(M)^*.$$

Taking norm on both sides and rearranging, using the fact that $\|\Im[M]\| \le \|M\|$ gives us the first inequality. Also,

$$\Im[z]\mathcal{F}(M)\mathcal{F}(M)^* \succeq \frac{\Im[z]}{\|\mathbb{E}H - \mathbb{E}[(H-\mathbb{E}H)M(H-\mathbb{E}H)] - zI_n\|^2}. \qquad \square$$

Since $\mathcal{F}$ is a holomorphic function mapping $\mathscr{M}_b$ strictly into itself, if follows from the Earle-Hamilton fixed point theorem that $\mathcal{F}$ is contractive with respect to the CRF-pseudometric ρ. Also, $\rho(X,Y) \ge c\|X-Y\|$. Hence, we can use the Banach fixed point theorem to show that $\mathcal{F}$ has a unique fixed point in $\mathscr{M}_b$ for every $b \in \mathbb{R}_{>0}$.

Theorem 1. *There exists a unique solution to the matrix Dyson equation in the admissible set.*

2.2 Stability

The next step is to show that the solution to the matrix Dyson equation is stable with respect to small additive perturbation. That is, say that F satisfies

$$(\mathbb{E}H - \mathbb{E}[(H - \mathbb{E}H)F(H - \mathbb{E}H)] - zI_n)F = I_n + D$$

where D is a perturbation matrix. If D is small, we want to show that F is close to the solution of the unperturbed matrix Dyson equation. This can be derived using the CRF-pseudometric.

Theorem 2. *If $\|D\| \to 0$ as $n \to \infty$, then $\|F(z) - M(z)\| \to 0$ as $n \to \infty$.*

2.3 Perturbation

Hence, if we are given a matrix H and we want to find a deterministic equivalent for the resolvent of H, we simply have to write

$$(\mathbb{E}H - \mathbb{E}_{\widetilde{H}}[(\widetilde{H} - \mathbb{E}H)(H - zI_n)^{-1}(\widetilde{H} - \mathbb{E}H)] - zI_n)(H - zI_n)^{-1} = I_n + D$$

and show that D is small in a suitable sense. This streamlines the process of finding deterministic equivalents for resolvents of random matrices.

3 Linearization Trick

The matrix Dyson equation is a powerful tool because it allows us to systematically find deterministic equivalents for resolvents of random matrices. However, it is limited to "Wigner-like" matrices. For instance, it does not give the right deterministic equivalent for Wishart matrices.

Example 1. Let $Z \in \mathbb{R}^{d\times n}$ be a matrix with i.i.d. standard Gaussian entries and define the Wishart matrix $W = n^{-1}ZZ^T \in \mathbb{R}^{d\times d}$. Let $M \in \mathbb{C}^{d\times d}$ be any deterministic matrix. Indeed, $\mathbb{E}W = I_d$ and

$$\mathbb{E}[(W - \mathbb{E}W)M(W - \mathbb{E}W)] = \frac{1}{n}\left(\mathbb{E}[(z^TMz)zz^T] - M\right).$$

To evaluate $\mathbb{E}[(z^TMz)zz^T]$, we use Wick's formula to compute

$$\mathbb{E}[(z^TMz)zz^T]_{j,k} = \sum_{a,b=1}^{d} M_{a,b}\mathbb{E}[z_az_bz_jz_k] = \sum_{a,b=1}^{d} M_{a,b}(\delta_{a,b}\delta_{j,k} + \delta_{a,j}\delta_{b,k} + \delta_{a,k}\delta_{b,j}),$$

which in matrix form can be written as $\mathbb{E}[(z^TMz)zz^T] = \operatorname{tr}(M)I_d + M + M^T$. Hence,

$$\mathbb{E}[(W - \mathbb{E}W)M(W - \mathbb{E}W)] = n^{-1}\operatorname{tr}(M)I_d + n^{-1}M^T.$$

For "reasonable" M, the term $n^{-1}M^T$ is negligible compared to the term $n^{-1}\operatorname{tr}(M)I_d$ and we can write

$$\mathbb{E}[(W-\mathbb{E}W)M(W-\mathbb{E}W)] = n^{-1}\operatorname{tr}(M)I_d + o_n(1).$$

Plugging this into the matrix Dyson equation and denoting $m = d^{-1}\operatorname{tr}(M)$, we get (to leading order) $m = (1-\frac{d}{n}m-z)^{-1}$, which simplifies to $m(1-\frac{d}{n}m-z)=1$. We can rearrange this to

$$\frac{d}{n}m^2 - (1-z)m + 1 = 0.$$

We know that this is wrong, since the true equation for the Stieltjes transform of the Wishart matrix is

$$z\frac{d}{n}m^2 - (1-z-\frac{d}{n})m + 1 = 0.$$

In fact, $\frac{d}{n}m^2-(1-z)m+1=0$ is precisely the equation for the Stieltjes transform of a Wigner model with uniform entrywise variance d/n, subjected to an additive deformation by the identity matrix I_d which uniformly shifts the spectrum by 1.

This demonstrates that while the standard MDE perfectly captures generalized Wigner ensembles, it fails to recover the Marchenko-Pastur law because it does not account for the heavy entrywise correlations inherent in sample covariance matrices.

To overcome this limitation, we can use the linearization trick. The concept of using linearizations, which are also referred to as linear pencils or realizations, is to represent rational functions of random matrices as blocks of inverse of larger random matrices which depend linearly on their blocks. These linearizations possess simpler correlation structures, rendering them more amenable to certain types of analysis.

Proposition 3 ([And13])**.** *Let p be a self-adjoint n by n polynomial expression in $\mathbb{C}\langle X_1,\dots,X_k\rangle$. Then, there exists a self-adjoint linearization $L \in \mathbb{C}^{(n+d)\times(n+d)}$ such that*

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

Example 2 (Linearization for Wishart matrix)**.** Let $X \in \mathbb{C}^{n\times d}$ and $z \in \mathbb{H}$. Then,

$$\begin{bmatrix} -zI_n & X \\ X^* & -I_d \end{bmatrix}^{-1}_{1,1} = (XX^* - zI_n)^{-1}.$$

Example 3 (Linearization for sample covariance matrix)**.** Let $X \in \mathbb{C}^{n\times d}$, $Y \in \mathbb{C}^{d\times m}$ and $z \in \mathbb{H}$. Then,

$$\begin{bmatrix} -zI_n & 0 & 0 & X \\ 0 & 0 & Y & -I_d \\ 0 & Y^* & -I_m & 0 \\ X^* & -I_d & 0 & 0 \end{bmatrix}^{-1}_{1,1} = (XYY^*X^* - zI_n)^{-1}.$$

Example 4 (Linearization for the anticommutator)**.** Let $X, Y \in \mathbb{C}^{n\times d}$ and $z \in \mathbb{H}$. Then,

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

The linearization trick as been used thoroughly in the free probability literature.

4 Matrix Dyson Equation for Linearizations

The linearization trick naturally leads to the study of *pseudo-resolvents.*

Definition 4. Let $L \in \mathbb{C}^{(n+d)\times(n+d)}$ be a linearization and $z \in \mathbb{H}$. The pseudo-resolvent of L is defined by $(L - z\Lambda)^{-1}$ where $\Lambda = \operatorname{diag}\{I_n, 0_{d\times d}\}$.

We consider real linearizations of the form

$$L = \begin{bmatrix} A & B^T \\ B & Q \end{bmatrix}$$

where $A \in \mathbb{R}^{n\times n}$ is self-adjoint, $B \in \mathbb{R}^{d\times n}$ and $Q \in \mathbb{R}^{d\times d}$ is deterministic and invertible. The matrix Dyson equation for linearizations is given by

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

where

$$\mathcal{S} : M \in \mathbb{C}^{\ell\times\ell} \mapsto \begin{bmatrix} \mathbb{E}[(L - \mathbb{E}L)M(L - \mathbb{E}L)]_{1,1} & \mathbb{E}[(A - \mathbb{E}A)M_{1,1}(B - \mathbb{E}B)^T] \\ \mathbb{E}[(B - \mathbb{E}B)M_{1,1}(A - \mathbb{E}A)] & \mathbb{E}[(B - \mathbb{E}B)M_{1,1}(B - \mathbb{E}B)^T] \end{bmatrix} \in \mathbb{C}^{\ell\times\ell} \quad (1)$$

is the superoperator. The associated admissible set is given by

$$\mathscr{M} := \{f : \mathbb{H} \mapsto \mathscr{A} \text{ analytic}\}, \quad \mathscr{A} := \{M \in \mathbb{C}^{\ell\times\ell} : \Im[M] \succeq 0 \text{ and } \Im[M_{1,1}] \succ 0\}.$$

The main difficulty in this setting is that, since the spectral parameter z does not span the entire diagonal, we cannot directly use the same fixed point theory as for the matrix Dyson equation and we have generally less a priori control. For instance, it is not clear a priori that the MDE map $M \in \mathscr{A} \mapsto (\mathbb{E}L - \mathcal{S}(M) - z\Lambda)^{-1} \in \mathscr{A}$ is well-defined in this case.

To study the matrix Dyson equation for linearizations, we introduce the regularized matrix Dyson equation (RMDE)

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

We have existence, stability, and nice properties for the RMDE. We can then show that the RMDE converges to the MDE as $\tau \to 0$. This allows us to show existence and uniqueness of a solution to the MDE for linearizations.

Theorem 3 (L.-V., Paquette '23)**.** *There exists a unique analytic matrix-valued function $M \in \mathscr{M}$ such that $M(z)$ solves the MDE for every $z \in \mathbb{H}$. Additionally, $\|M_{1,1}(z)\| \leq (\Im[z])^{-1}$ and*

$$M(z) = \begin{bmatrix} 0_{n\times n} & 0_{n\times d} \\ 0_{d\times n} & Q^{-1} \end{bmatrix} + \int_{\mathbb{R}} \frac{\Omega(\mathrm{d}\lambda)}{\lambda - z}$$

for all $z \in \mathbb{H}$, where Ω is a real Borel $\ell \times \ell$ positive semidefinite compactly supported measure satisfying

$$\int_{\mathbb{R}} \Omega(\mathrm{d}\lambda) = \begin{bmatrix} I_n & -\mathbb{E}[B^T]Q^{-1} \\ -Q^{-1}\mathbb{E}[B] & Q^{-1}\mathbb{E}[BB^T]Q^{-1}. \end{bmatrix}.$$

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

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

We have the following stability result.

Theorem 4 (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* $z \in \mathbb{H}$.

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

Let us discuss the assumptions in the last theorem. First, the assumption $\|M^{(\tau)}(z) - M(z)\| \xrightarrow{\tau \to 0} 0$ uniformly in ℓ ensures the stability of the MDE. When the linearization contains block of independent Wigner for instance, this assumption is satisfied because one can construct a dimension independent representation of the solution of the (R)MDE.

For the third assumption, we have the following result.

Theorem 5 (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.*

Combining everything with a concentration argument, we can show that the unique solution to the MDE for linearizations is a deterministic equivalent for the pseudo-resolvent.

Theorem 6. *Let $L = \mathcal{C}(g) + \mathbb{E}L$ for some $g \in \mathcal{N}(0, I_\gamma)$. Under some technical assumptions, $\operatorname{tr}(U((L - z\Lambda)^{-1} - M(z))) \xrightarrow[\ell \to \infty]{a.s.} 0$ for every $U \in \mathbb{C}^{\ell \times \ell}$ with $\|U\|_* \leq 1$.*

5 Application

Consider the random features ridge regression problem:

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

where

- $A = n^{-\frac{1}{2}}\sigma(XW)$
- $X \in \mathbb{R}^{n_{\text{train}} \times n_0}$ be the matrix with jth rows corresponding to x_j^T (the jth sample)
- y be the vectors of labels
- $W \in \mathbb{R}^{n_0 \times d}$ is a matrix of i.i.d. Gaussians
- σ Lipschitz functions applied entrywise
- ridge parameter $\delta > 0$
- $\mathbb{E}A = 0$

The random features model is equivalent to forming a two-layer neural network and fixing the weights of the first layer at random initialization. The random features model has been studied to understand things like implicit regularization, multiple descent, universality, etc. that have been observed in practice.

The problem above admits the explicit solution $w_{\text{ridge}} = A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}y$. Given another labeled dataset $\widehat{\mathscr{D}} = \{(\widehat{x}_j, \widehat{y}_j)\}_{j=1}^{n_{\text{test}}}$, we can compute the empirical test error, or out-of-sample error, using the squared norm of the residuals

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

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

Theorem 7. *Suppose that* $n_{\text{train}}, d, n_{\text{test}}, n_0 \propto n$ *and* $\limsup_{n\to\infty} \max\{\mathbb{E}[\|A\|^4], \mathbb{E}[\|\widehat{A}\|^4]\} < \infty$. *Let* α *be the unique non-positive real number satisfying*

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

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

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

Then, $d\beta\|K_{AA^T}^{\frac{1}{2}} My\|^2 + \|d\alpha K_{\widehat{A}A^T} My + \widehat{y}\|^2 - E_{\text{test}} \xrightarrow[n\to\infty]{a.s.} 0$.

Here, we assumed that

$$\mathbb{E}[(a_1^T, \hat{a}_1^T)^T] = 0 \quad \text{and} \quad \mathbb{E}[(a_1^T, \hat{a}_1^T)^T(a_1^T, \hat{a}_1^T)] = \begin{bmatrix} K_{AA^T} & K_{A\widehat{A}^T} \\ K_{\widehat{A}A^T} & K_{\widehat{A}\widehat{A}^T} \end{bmatrix}.$$

We start by considering $\widehat{A}A^T\left(AA^T + \delta I_{n_{\text{train}}}\right)^{-1}$. Let $\ell = n_{\text{train}} + d + 2n_{\text{test}}$ and consider the linearization

$$L - z\Lambda = \begin{bmatrix} (\delta - z)I_{n_{\text{train}}} & A & 0_{n_{\text{train}} \times n_{\text{test}}} & 0_{n_{\text{train}} \times n_{\text{test}}} \\ A^T & -(1+z)I_{d\times d} & 0_{d \times n_{\text{test}}} & \widehat{A}^T \\ 0_{n_{\text{test}} \times n_{\text{train}}} & 0_{n_{\text{test}} \times d} & 0_{n_{\text{test}} \times n_{\text{test}}} & -I_{n_{\text{test}}} \\ 0_{n_{\text{test}} \times n_{\text{train}}} & \widehat{A} & -I_{n_{\text{test}}} & 0_{n_{\text{test}} \times n_{\text{test}}} \end{bmatrix} \in \mathbb{C}^{\ell \times \ell}. \tag{2}$$

Then

$$(L - z\Lambda)^{-1} = \begin{bmatrix} R & (1+z)^{-1}RA & (1+z)^{-1}RA\widehat{A}^T & 0 \\ (1+z)^{-1}A^TR & \bar{R} & \bar{R}\widehat{A}^T & 0 \\ (1+z)^{-1}\widehat{A}A^TR & \widehat{A}\bar{R} & \widehat{A}\bar{R}\widehat{A}^T & -I_{n_{\text{test}}} \\ 0 & 0 & -I_{n_{\text{test}}} & 0 \end{bmatrix}.$$

Here $R := ((1+z)^{-1}AA^T + (\delta - z)I_{n_{\text{train}}})^{-1}$ represents a resolvent and $\bar{R} := -((1+z)I_d + (\delta - z)^{-1}A^TA)^{-1}$ is a co-resolvent. Indeed, the term $\lim_{z\to 0}(L - z\Lambda)^{-1}_{3,1} = \widehat{A}A^T\left(AA^T + \delta I_{n_{\text{train}}}\right)^{-1}$ holds the relevant expression. Therefore, it suffices to find a deterministic equivalent for the pseudo-resolvent $(L - z\Lambda)^{-1}$ and take the spectral parameter to zero.

The associated MDE is given by

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

Using our framework, we can show that the solution to this equation is a deterministic equivalent for the pseudo-resolvent. We can then take the spectral parameter to zero and extract the (3,1) block of interest.

Next, we want to find a deterministic equivalent for $(AA^T + \delta I_{n_{\text{train}}})^{-1}A\widehat{A}^T\widehat{A}A^T(AA^T + \delta I_{n_{\text{train}}})^{-1}$. This is the square of the previous matrix, so we can use a contour integral trick to find a deterministic equivalent for this expression.

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.

[Bel10] Richard E. Bellman. *Dynamic Programming*. Princeton: Princeton University Press, 2010. ISBN: 9781400835386. DOI: doi:10.1515/9781400835386.

[Bro+20] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[CL22] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[Ram+21] Aditya Ramesh et al. "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 8821–8831.

[Ram+22] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022.

[Zha+21] Chiyuan Zhang et al. "Understanding deep learning (still) requires rethinking generalization". In: *Commun. ACM* 64.3 (Feb. 2021), pp. 107–115. ISSN: 0001-0782. DOI: 10.1145/3446776.