Applications of random matrix theory to graph matching and neural networks

Zhou Fan

Department of Statistics and Data Science
Yale University

(Online) Random Matrices and Their Applications 2020

Outline

In this talk, I'll discuss applications of random matrix theory to two (unrelated) problems in statistics and machine learning:

- Graph matching
- Spectral analysis of neural network kernel matrices

Outline

In this talk, I'll discuss applications of random matrix theory to two (unrelated) problems in statistics and machine learning:

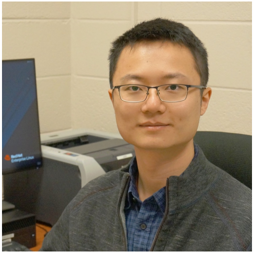
- Graph matching
- Spectral analysis of neural network kernel matrices

I'll focus on high-level ideas, discuss the random matrix connections, and describe a few open questions.

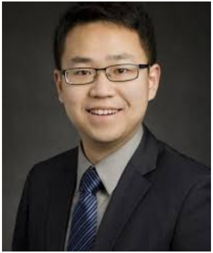
Graph Matching

Graph matching

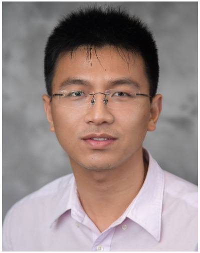
Joint work with:



Cheng Mao

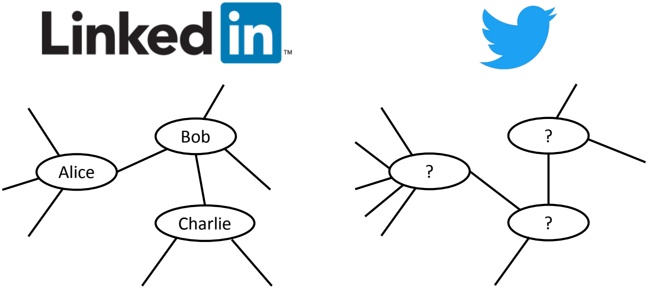


Yihong Wu



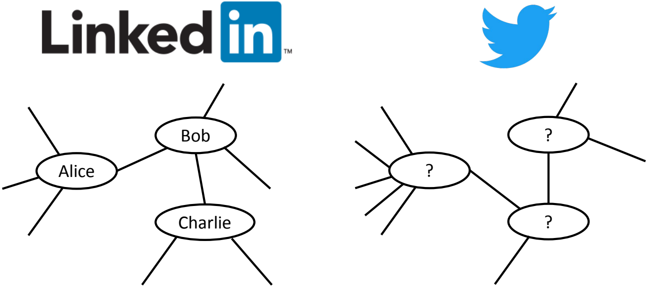
Jiaming Xu

Graph matching



[Picture courtesy of R. Srikant]

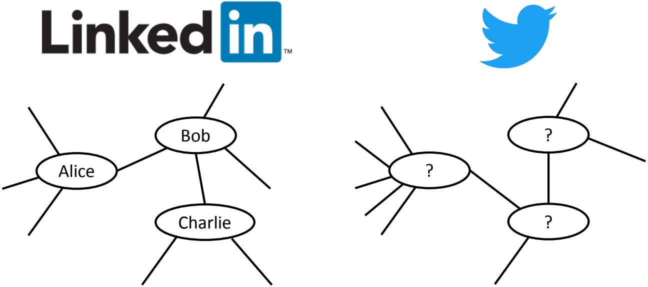
Graph matching



[Picture courtesy of R. Srikant]

Given the LinkedIn network, can you de-anonymize Twitter?

Graph matching

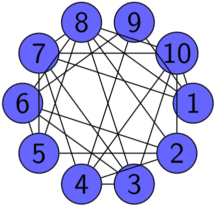


[Picture courtesy of R. Srikant]

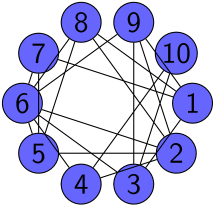
Given the LinkedIn network, can you de-anonymize Twitter?

More abstractly: Given two *correlated* random graphs on n vertices, with a hidden correspondence between their vertices, can you recover this vertex matching?

Correlated Erdős-Rényi graph model



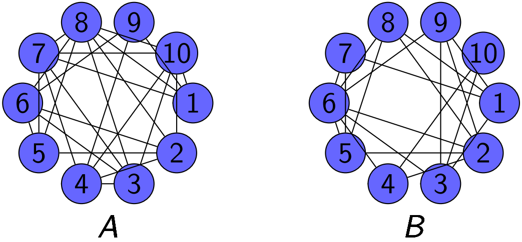
A



B

$$A_{ij}, B_{ij} \sim \text{Bernoulli}(q) \quad \text{and} \quad \mathbb{P}[A_{ij} = B_{ij} = 1] = (1-\delta)q$$

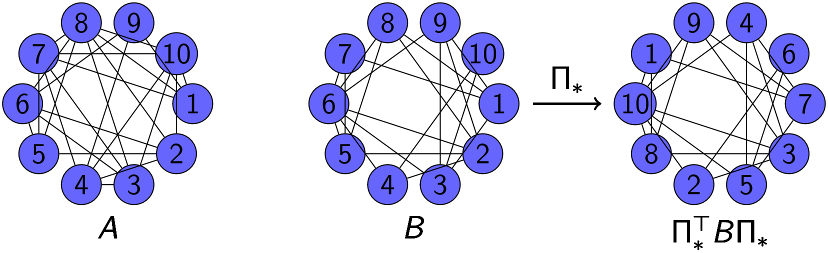
Correlated Erdős-Rényi graph model



$$A_{ij}, B_{ij} \sim \text{Bernoulli}(q) \quad \text{and} \quad \mathbb{P}[A_{ij} = B_{ij} = 1] = (1 - \delta)q$$

q is the sparsity, and δ is the fraction of differing edges.
Different edge pairs (i, j) are independent. [Pedarsani, Grossglauser '11]

Correlated Erdős-Rényi graph model

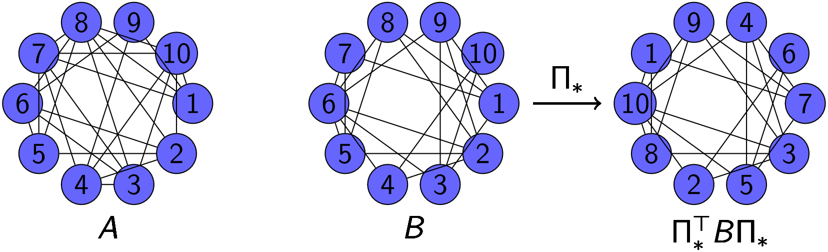


$$A_{ij}, B_{ij} \sim \text{Bernoulli}(q) \quad \text{and} \quad \mathbb{P}[A_{ij} = B_{ij} = 1] = (1-\delta)q$$

q is the sparsity, and δ is the fraction of differing edges.
Different edge pairs (i, j) are independent. [Pedarsani, Grossglauser '11]

We observe A and $\Pi_*^\top B \Pi_*$ and want to recover Π_*.

Correlated Erdős-Rényi graph model



$$A_{ij}, B_{ij} \sim \text{Bernoulli}(q) \quad \text{and} \quad \mathbb{P}[A_{ij} = B_{ij} = 1] = (1-\delta)q$$

q is the sparsity, and δ is the fraction of differing edges.
Different edge pairs (i, j) are independent. [Pedarsani, Grossglauser '11]

We observe A and $\Pi_*^\top B \Pi_*$ and want to recover Π_*. Questions:

- How correlated must A and B be, to recover Π_* w.h.p.?
- How to design a computational algorithm that achieves this?

Spectral algorithms

Use the (permutation invariant) eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

Spectral algorithms

Use the (permutation invariant) eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

- **Top eigenvector:** Match A to B by sorting u_1 and v_1. Similar ideas in IsoRank [Singh, Xu, Berger '08], EigenAlign [Feizi et al '19].

Spectral algorithms

Use the (permutation invariant) eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

- **Top eigenvector:** Match A to B by sorting u_1 and v_1. Similar ideas in IsoRank [Singh, Xu, Berger '08], EigenAlign [Feizi et al '19].
- **All eigenvectors:** Find the permutation Π which maximizes

$$\sum_{i=1}^{n} v_i^\top \Pi u_i \equiv \operatorname{Tr} X\Pi \quad \text{where} \quad X = \sum_{i=1}^{n} u_i v_i^\top$$

This aligns every u_i with the corresponding v_i. [Umeyama '88]

Spectral algorithms

Use the (permutation invariant) eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

- **Top eigenvector:** Match A to B by sorting u_1 and v_1. Similar ideas in IsoRank [Singh, Xu, Berger '08], EigenAlign [Feizi et al '19].
- **All eigenvectors:** Find the permutation Π which maximizes

$$\sum_{i=1}^{n} v_i^\top \Pi u_i \equiv \operatorname{Tr} X\Pi \quad \text{where} \quad X = \sum_{i=1}^{n} u_i v_i^\top$$

  This aligns every u_i with the corresponding v_i. [Umeyama '88]

Both work in noiseless settings ($\delta = 0$), but are brittle to noise: Each pair (u_i, v_i) decorrelates when $\delta > 1/n^\alpha$ for some $\alpha > 0$.

A new spectral algorithm: GRAMPA

GRAph Matching by Pairwise eigen-Alignments

1. Compute the eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

A new spectral algorithm: GRAMPA

GRAph Matching by Pairwise eigen-Alignments

1. Compute the eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

2. Construct the similarity matrix

$$X = \sum_{i,j=1}^{n} \underbrace{\frac{\eta}{(\lambda_i - \mu_j)^2 + \eta^2}}_{\text{Cauchy kernel applied to } \lambda_i \text{ and } \mu_j} \times \underbrace{u_i u_i^\top \mathbf{J} v_j v_j^\top}_{\text{“Alignment” between } u_i \text{ and } v_j}$$

where $\eta =$ bandwidth parameter, $\mathbf{J} =$ all-1's matrix.

A new spectral algorithm: GRAMPA

GRAph Matching by Pairwise eigen-Alignments

1. Compute the eigendecompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \quad \text{and} \quad B = \sum_{j=1}^{n} \mu_j v_j v_j^\top$$

2. Construct the similarity matrix

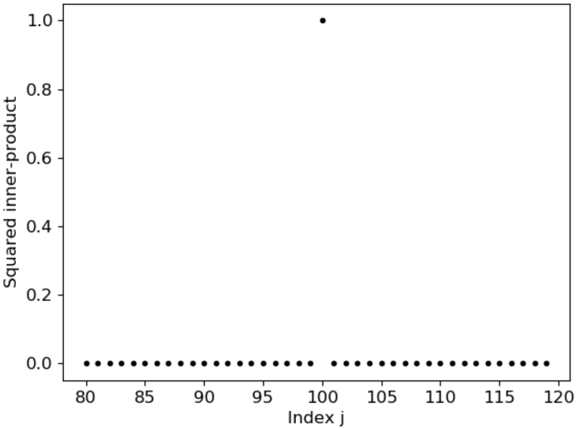
$$X = \sum_{i,j=1}^{n} \underbrace{\frac{\eta}{(\lambda_i - \mu_j)^2 + \eta^2}}_{\text{Cauchy kernel applied to } \lambda_i \text{ and } \mu_j} \times \underbrace{u_i u_i^\top \mathbf{J} v_j v_j^\top}_{\text{“Alignment” between } u_i \text{ and } v_j}$$

   where $\eta =$ bandwidth parameter, $\mathbf{J} =$ all-1's matrix.

3. Find the permutation Π which maximizes $\operatorname{Tr} X\Pi$. This tries to align every u_i with every v_j, with weighting by the Cauchy kernel.

Motivation for GRAMPA

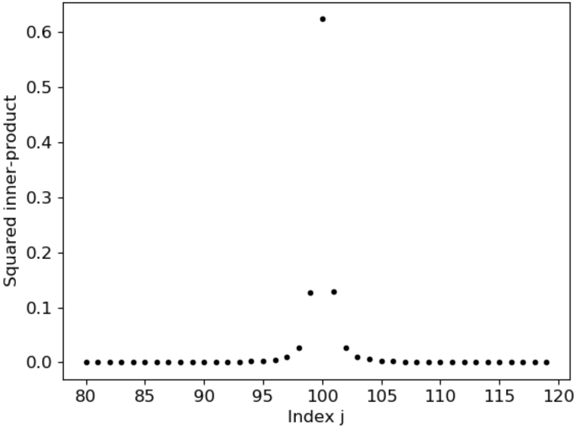
Isomorphic Erdős-Rényi graphs (500 vertices, edge probability $\frac{1}{2}$)



$\langle u_{100}, v_j \rangle^2$ for $j \in \{80, \ldots, 120\}$, averaged across 1000 simulations

Motivation for GRAMPA

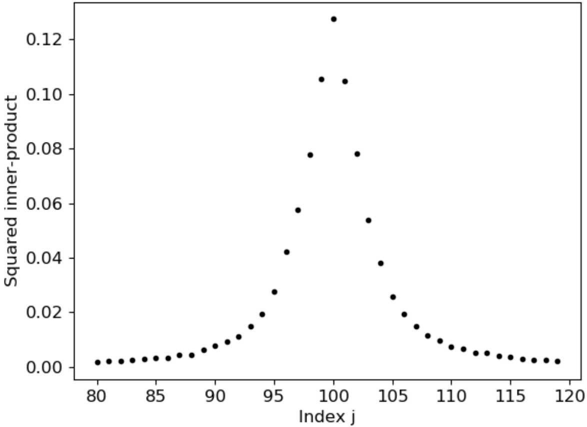
Erdős-Rényi graphs with fraction of differing edges $\delta = 0.001$



$\langle u_{100}, v_j \rangle^2$ for $j \in \{80, \ldots, 120\}$, averaged across 1000 simulations

Motivation for GRAMPA

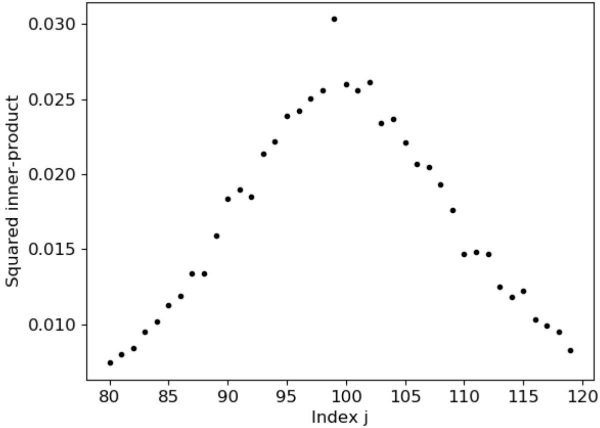
Erdős-Rényi graphs with fraction of differing edges $\delta = 0.01$



$\langle u_{100}, v_j \rangle^2$ for $j \in \{80, \ldots, 120\}$, averaged across 1000 simulations

Motivation for GRAMPA

Erdős-Rényi graphs with fraction of differing edges $\delta = 0.05$



$\langle u_{100}, v_j \rangle^2$ for $j \in \{80, \ldots, 120\}$, averaged across 1000 simulations

Motivation for GRAMPA

$$X = \sum_{i,j=1}^{n} \underbrace{\frac{\eta}{(\lambda_i - \mu_j)^2 + \eta^2}}_{\text{Cauchy kernel applied to } \lambda_i \text{ and } \mu_j} \times \underbrace{u_i u_i^\top \mathbf{J} v_j v_j^\top}_{\text{“Alignment” between } u_i \text{ and } v_j}$$

Motivation for GRAMPA

$$X = \sum_{i,j=1}^{n} \underbrace{\frac{\eta}{(\lambda_i - \mu_j)^2 + \eta^2}}_{\text{Cauchy kernel applied to } \lambda_i \text{ and } \mu_j} \times \underbrace{u_i u_i^\top \mathbf{J} v_j v_j^\top}_{\text{“Alignment” between } u_i \text{ and } v_j}$$

The Cauchy kernel may be motivated by eigenvector correlation decay in the Dyson Brownian motion model

$$B = A + Z_\delta$$

where $Z \stackrel{L}{=} \sqrt{\delta} \times$ independent GOE. Results of [Benigni '17] show, using analysis of the eigenvector moment flow in [Bourgade, Yau '17], that

$$n \cdot \mathbb{E}[\langle u_i, v_j \rangle^2] \approx \frac{\delta}{(\lambda_i - \mu_j)^2 + C\delta^2}$$

Motivation for GRAMPA

$$X = \sum_{i,j=1}^{n} \underbrace{\frac{\eta}{(\lambda_i - \mu_j)^2 + \eta^2}}_{\text{Cauchy kernel applied to } \lambda_i \text{ and } \mu_j} \times \underbrace{u_i u_i^\top \mathbf{J} v_j v_j^\top}_{\text{"Alignment" between } u_i \text{ and } v_j}$$

The Cauchy kernel may be motivated by eigenvector correlation decay in the Dyson Brownian motion model

$$B = A + Z_\delta$$

where $Z \stackrel{L}{=} \sqrt{\delta} \times$ independent GOE. Results of [Benigni '17] show, using analysis of the eigenvector moment flow in [Bourgade, Yau '17], that

$$n \cdot \mathbb{E}[\langle u_i, v_j \rangle^2] \approx \frac{\delta}{(\lambda_i - \mu_j)^2 + C\delta^2}$$

[Question: Is this true also for a time-evolving Erdős-Rényi model?]

Theoretical guarantee

Theorem (F., Mao, Wu, Xu)

For the correlated Erdős-Rényi model with edge probability $q \geq \operatorname{polylog}(n)/n$ *and fraction of differing edges* $\delta \leq 1/\operatorname{polylog}(n)$, *this algorithm recovers the true vertex correspondence* Π_* *w.h.p.*

Theoretical guarantee

Theorem (F., Mao, Wu, Xu)

For the correlated Erdős-Rényi model with edge probability $q \geq \mathrm{polylog}(n)/n$ *and fraction of differing edges* $\delta \leq 1/\mathrm{polylog}(n)$, *this algorithm recovers the true vertex correspondence* Π_* *w.h.p.*

- Improves over previous spectral algorithms requiring $\delta \leq 1/n^{\alpha}$.

Theoretical guarantee

Theorem (F., Mao, Wu, Xu)

For the correlated Erdős-Rényi model with edge probability $q \geq \operatorname{polylog}(n)/n$ and fraction of differing edges $\delta \leq 1/\operatorname{polylog}(n)$, this algorithm recovers the true vertex correspondence Π_* w.h.p.

- Improves over previous spectral algorithms requiring $\delta \leq 1/n^{\alpha}$.
- This is currently the best-known guarantee for polynomial-time algorithms. Matches previous result of [Ding, Ma, Wu, Xu '18].

Theoretical guarantee

Theorem (F., Mao, Wu, Xu)

For the correlated Erdős-Rényi model with edge probability $q \geq \operatorname{polylog}(n)/n$ *and fraction of differing edges* $\delta \leq 1/\operatorname{polylog}(n)$, *this algorithm recovers the true vertex correspondence* Π_* *w.h.p.*

- Improves over previous spectral algorithms requiring $\delta \leq 1/n^\alpha$.
- This is currently the best-known guarantee for polynomial-time algorithms. Matches previous result of [Ding, Ma, Wu, Xu '18].
- Recovery of Π^* is possible once $\delta \leq 1 - 1/\operatorname{polylog}(n)$ [Cullina, Kiyavash '18], but no efficient algorithm is known.

Theoretical guarantee

Theorem (F., Mao, Wu, Xu)

For the correlated Erdős-Rényi model with edge probability $q \geq \mathrm{polylog}(n)/n$ and fraction of differing edges $\delta \leq 1/\mathrm{polylog}(n)$, this algorithm recovers the true vertex correspondence Π_* w.h.p.

- Improves over previous spectral algorithms requiring $\delta \leq 1/n^{\alpha}$.
- This is currently the best-known guarantee for polynomial-time algorithms. Matches previous result of [Ding, Ma, Wu, Xu '18].
- Recovery of Π^* is possible once $\delta \leq 1 - 1/\mathrm{polylog}(n)$ [Cullina, Kiyavash '18], but no efficient algorithm is known.
- [Barak, Chou, Lei, Schramm, Sheng '18] developed an $n^{O(\log n)}$-time algorithm, which succeeds for $\delta \leq 1 - \varepsilon$ and $q \geq n^{\varepsilon}/n$.
- [Ganassali, Massoulié '20] developed a polynomial-time algorithm that recovers a positive fraction of the vertex matchings, for $\delta \leq 1 - c$ and $q \asymp 1/n$.

Main ideas of the analysis

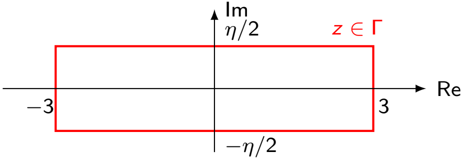
Define the resolvents

$$R_A(z) = (A - z\,\mathrm{Id})^{-1} \qquad R_B(z) = (B - z\,\mathrm{Id})^{-1}$$

Lemma

The GRAMPA similarity matrix X has the resolvent representation

$$X = \frac{1}{2\pi}\,\mathrm{Re}\oint_\Gamma R_A(z)\mathbf{J}R_B(z+\mathbf{i}\eta)dz$$



This contour Γ contains all of the poles of R_A, and none of the poles of R_B.

Main ideas of the analysis

Suppose $\Pi^* = \mathrm{Id}$, and consider the (k, ℓ) entry

$$X_{k\ell} = \frac{1}{2\pi} \operatorname{Re} \oint_{\Gamma} \left[e_k^{\top} R_A(z) \mathbf{J} R_B(z + \mathbf{i}\eta) e_\ell \right] dz$$

Main ideas of the analysis

Suppose $\Pi^* = \mathrm{Id}$, and consider the (k,ℓ) entry

$$X_{k\ell} = \frac{1}{2\pi} \operatorname{Re} \oint_\Gamma \left[e_k^\top R_A(z) \mathbf{J} R_B(z + \mathbf{i}\eta) e_\ell \right] dz$$

Diagonal: By Schur-complement identities,

$$X_{kk} \approx \frac{1}{2\pi} \operatorname{Re} a_k^\top \left[\oint_\Gamma m(z) m(z + \mathbf{i}\eta) R_{A^{(k)}}(z) \mathbf{J} R_{B^{(k)}}(z + \mathbf{i}\eta)\, dz \right] b_k$$

(a_k, b_k) in (A, B) are correlated, and independent of $(A^{(k)}, B^{(k)})$.

Main ideas of the analysis

Suppose $\Pi^* = \mathrm{Id}$, and consider the (k, ℓ) entry

$$X_{k\ell} = \frac{1}{2\pi} \operatorname{Re} \oint_\Gamma \left[e_k^\top R_A(z) \mathbf{J} R_B(z + \mathbf{i}\eta) e_\ell \right] dz$$

Diagonal: By Schur-complement identities,

$$X_{kk} \approx \frac{1}{2\pi} \operatorname{Re} a_k^\top \left[\oint_\Gamma m(z) m(z + \mathbf{i}\eta) R_{A^{(k)}}(z) \mathbf{J} R_{B^{(k)}}(z + \mathbf{i}\eta)\, dz \right] b_k$$

(a_k, b_k) in (A, B) are correlated, and independent of $(A^{(k)}, B^{(k)})$.

Off-diagonal: Similarly,

$$X_{k\ell} \approx \frac{1}{2\pi} \operatorname{Re} a_k^\top \left[\oint_\Gamma m(z) m(z + \mathbf{i}\eta) R_{A^{(k\ell)}}(z) \mathbf{J} R_{B^{(k\ell)}}(z + \mathbf{i}\eta) dz \right] b_\ell$$

(a_k, b_ℓ) are independent, and also independent of $(A^{(k\ell)}, B^{(k\ell)})$.

Main ideas of the analysis

Applying local law estimates and fluctuation averaging techniques from [Erdős, Knowles, Yau, Yin '13], we analyze the traces and Frobenius norms of the preceding integrals.

Main ideas of the analysis

Applying local law estimates and fluctuation averaging techniques from [Erdős, Knowles, Yau, Yin '13], we analyze the traces and Frobenius norms of the preceding integrals.

When $\Pi^* = \mathrm{Id}$,

$$\min_k X_{kk} > \max_{k \neq \ell} X_{k\ell} \quad \text{w.h.p.}$$

Then the permutation Π maximizing $\operatorname{Tr} X\Pi$ is $\Pi = \mathrm{Id}$, so GRAMPA returns Id w.h.p.

By permutation invariance of the algorithm, GRAMPA returns Π_* w.h.p. for any true permutation Π^*.

A different motivation for GRAMPA

$$\min_{\Pi \in S_n} \|A - \Pi^\top B \Pi\|_F^2 = \min_{\Pi \in S_n} \|\Pi A - B\Pi\|_F^2$$

A different motivation for GRAMPA

$$\min_{\Pi \in S_n} \|A - \Pi^\top B \Pi\|_F^2 = \min_{\Pi \in S_n} \|\Pi A - B \Pi\|_F^2$$

Relax this to the quadratic program

$$\min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2$$

for the convex hull $\mathrm{conv}(S_n) = \{X : X_{ij} \geq 0,\ X1 = 1,\ X^\top 1 = 1\}$.

A different motivation for GRAMPA

$$\min_{\Pi \in S_n} \|A - \Pi^\top B \Pi\|_F^2 = \min_{\Pi \in S_n} \|\Pi A - B\Pi\|_F^2$$

Relax this to the quadratic program

$$\min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2$$

for the convex hull $\mathrm{conv}(S_n) = \{X : X_{ij} \geq 0,\ X1 = 1,\ X^\top 1 = 1\}$.

Solve this for X, then round to a permutation Π.

[Zaslavskiy, Bach, Vert '09], [Aflalo, Bronstein, Kimmel '15]

A different motivation for GRAMPA

$$\min_{\Pi \in S_n} \|A - \Pi^\top B \Pi\|_F^2 = \min_{\Pi \in S_n} \|\Pi A - B\Pi\|_F^2$$

Relax this to the quadratic program

$$\min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2$$

for the convex hull $\mathrm{conv}(S_n) = \{X : X_{ij} \geq 0,\ X1 = 1,\ X^\top 1 = 1\}$.
Solve this for X, then round to a permutation Π.
[Zaslavskiy, Bach, Vert '09], [Aflalo, Bronstein, Kimmel '15]

This method is not well-understood for the Erdős-Rényi model.
The GRAMPA matrix X is, instead, the further relaxation

$$\min_{X :\ 1^\top X 1 = n} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2$$

A hierarchy of relaxations

$$
\begin{tikzcd}
 & \text{Full QP: } \min\limits_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2 \arrow[dl] \arrow[dr] & \\
\min\limits_{X:\, X\mathbf{1}=\mathbf{1},\, X^\top \mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2 \arrow[d] & & \min\limits_{X:\, X_{ij} \geq 0,\, \|X\|_F^2 = 1} \|XA - BX\|_F^2 \arrow[ddl] \\
\min\limits_{X:\, X\mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2 \arrow[dr] & & \\
 & \text{GRAMPA: } \min\limits_{X:\, \mathbf{1}^\top X \mathbf{1} = n} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2 &
\end{tikzcd}
$$

A hierarchy of relaxations

Full QP: $$\min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2$$

$$\min_{X:\, X\mathbf{1}=\mathbf{1},\, X^\top \mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2$$

$$\min_{X:\, X_{ij} \geq 0,\, \|X\|_F^2 = 1} \|XA - BX\|_F^2$$

$$\boxed{\min_{X:\, X\mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2}$$

$$\boxed{\text{GRAMPA:} \quad \min_{X:\, \mathbf{1}^\top X \mathbf{1} = n} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2}$$

These two relaxations have representations in terms of the spectra of A and B, and we analyze them in our work.

A hierarchy of relaxations

$$
\begin{tikzcd}[column sep=small]
 & \text{Full QP:} \quad \min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2 \arrow[dl] \arrow[dr] & \\
\boxed{\min_{X:\, X\mathbf{1}=\mathbf{1},\, X^\top \mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2} \arrow[d] & & \min_{X:\, X_{ij}\geq 0,\, \|X\|_F^2=1} \|XA - BX\|_F^2 \arrow[ddl] \\
\min_{X:\, X\mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2 \arrow[dr] & & \\
 & \text{GRAMPA:} \quad \min_{X:\, \mathbf{1}^\top X \mathbf{1}=n} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2 &
\end{tikzcd}
$$

Variants of this are related to the resolvent-type matrix

$$\Big[(A \otimes \mathrm{Id} - \mathrm{Id} \otimes B)^2 + \eta^2 (\mathbf{J} \otimes \mathrm{Id} + \mathrm{Id} \otimes \mathbf{J})\Big]^{-1}$$

for the Kronecker model $A \otimes \mathrm{Id} - \mathrm{Id} \otimes B \in \mathbb{R}^{n^2 \times n^2}$.

A hierarchy of relaxations

$$\boxed{\text{Full QP:} \quad \min_{X \in \mathrm{conv}(S_n)} \|XA - BX\|_F^2}$$

$$\min_{X:\, X\mathbf{1}=\mathbf{1},\, X^\top \mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2$$

$$\boxed{\min_{X:\, X_{ij}\geq 0,\, \|X\|_F^2=1} \|XA - BX\|_F^2}$$

$$\min_{X:\, X\mathbf{1}=\mathbf{1}} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2$$

$$\text{GRAMPA:} \quad \min_{X:\, \mathbf{1}^\top X \mathbf{1}=n} \|XA - BX\|_F^2 + \eta^2 \|X\|_F^2$$

How to analyze these programs with entrywise non-negativity is open. We believe from simulation that these may achieve exact recovery of Π^* w.h.p. up to $\delta \leq c$ for some constant $c > 0$.

Neural network kernel matrices

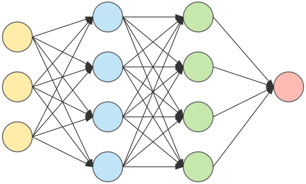
Neural network kernel matrices

Joint work with Zhichao Wang:

Feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined iteratively by

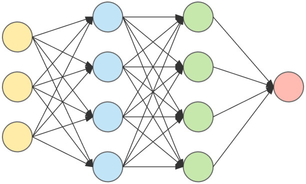
$$\mathbf{x}^1 = \sigma(W_1\mathbf{x}),\ \mathbf{x}^2 = \sigma(W_2\mathbf{x}^1),\ \ldots,\ \mathbf{x}^L = \sigma(W_L\mathbf{x}^{L-1}),\ f_\theta(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}^L$$



Feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined iteratively by

$$\mathbf{x}^1 = \sigma(W_1\mathbf{x}),\ \mathbf{x}^2 = \sigma(W_2\mathbf{x}^1),\ \ldots,\ \mathbf{x}^L = \sigma(W_L\mathbf{x}^{L-1}),\ f_\theta(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}^L$$

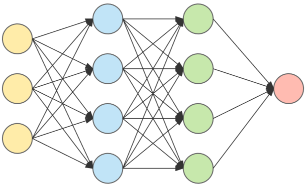


- $W_1 \in \mathbb{R}^{d_1\times d_0}$, $W_2 \in \mathbb{R}^{d_2\times d_1}$, ..., $W_L \in \mathbb{R}^{d_L\times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$ are the weights. We denote $\theta = (W_1, \ldots, W_L, \mathbf{w})$.
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, applied entrywise.

Feedforward neural network

Function $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$, $\mathbf{x} \mapsto f_\theta(\mathbf{x})$, defined iteratively by

$$\mathbf{x}^1 = \sigma(W_1\mathbf{x}),\ \mathbf{x}^2 = \sigma(W_2\mathbf{x}^1),\ \ldots,\ \mathbf{x}^L = \sigma(W_L\mathbf{x}^{L-1}),\ f_\theta(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}^L$$



- $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, ..., $W_L \in \mathbb{R}^{d_L \times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$ are the weights. We denote $\theta = (W_1, \ldots, W_L, \mathbf{w})$.
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function, applied entrywise.

Two fundamental questions:

- How does learning occur during gradient descent training of θ?
- What allows f_θ to generalize to unseen test samples?

Two kernel matrices

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be the training samples, and $X_\ell \in \mathbb{R}^{d_\ell \times n}$ the outputs of each layer $\ell = 1, \ldots, L$.

Recent theory of neural networks highlights two kernel matrices:

1. The **Conjugate Kernel** (or equivalent Gaussian process kernel)
$$K^{\mathsf{CK}} = X_L^\top X_L \in \mathbb{R}^{n \times n}$$

Two kernel matrices

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be the training samples, and $X_\ell \in \mathbb{R}^{d_\ell \times n}$ the outputs of each layer $\ell = 1, \ldots, L$.

Recent theory of neural networks highlights two kernel matrices:

1. The **Conjugate Kernel** (or equivalent Gaussian process kernel)

$$K^{\mathrm{CK}} = X_L^\top X_L \in \mathbb{R}^{n \times n}$$

   The final step of the network is just linear regression on X_L. K^{CK} governs the properties of this linear regression.

Two kernel matrices

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$ be the training samples, and $X_\ell \in \mathbb{R}^{d_\ell \times n}$ the outputs of each layer $\ell = 1, \ldots, L$.

Recent theory of neural networks highlights two kernel matrices:

1. The **Conjugate Kernel** (or equivalent Gaussian process kernel)

   $$K^{\mathrm{CK}} = X_L^\top X_L \in \mathbb{R}^{n \times n}$$

   The final step of the network is just linear regression on X_L. K^{CK} governs the properties of this linear regression.

   - The network is often already predictive when X_L is fixed by random initialization of $W_1, \ldots, W_L$, and only $\mathbf{w}$ is trained.
   - For $d_1, \ldots, d_L \to \infty$ and fixed n, K^{CK} converges to a limit kernel, and this is an approximation of regression in an associated RKHS.

   [Neal '94], [Williams '97], [Cho, Saul '09], [Rahimi, Recht '09], [Daniely et al '16], [Poole et al '16], [Schoenholz et al '17], [Lee et al '18], ...

Two kernel matrices

2. The **Neural Tangent Kernel**

$$K^{\mathsf{NTK}} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n \times n}$$

Two kernel matrices

2. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n \times n}$$

Training errors evolve during gradient descent as

$$\frac{d}{dt}\Big(\mathbf{y} - f_{\theta(t)}(X)\Big) = -K^{\mathrm{NTK}}(t) \cdot \Big(\mathbf{y} - f_{\theta(t)}(X)\Big)$$

Two kernel matrices

2. The **Neural Tangent Kernel**

$$K^{\text{NTK}} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n \times n}$$

Training errors evolve during gradient descent as

$$\frac{d}{dt}\Big(\mathbf{y} - f_{\theta(t)}(X)\Big) = -K^{\text{NTK}}(t) \cdot \Big(\mathbf{y} - f_{\theta(t)}(X)\Big)$$

- For $d_1, \ldots, d_L \to \infty$ and fixed n, K^{NTK} is constant over training.
- Then (diagonalizing K^{NTK}) $\mathbf{y} - f_{\theta(t)}(X) \to 0$ at a different exponential rate along each eigenvector of K^{NTK}.

[Jacot, Gabriel, Hongler '19], [Du et al '19], [Allen-Zhu et al '19], [Lee et al '19], ...

Two kernel matrices

2. The **Neural Tangent Kernel**

$$K^{\mathsf{NTK}} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n \times n}$$

Training errors evolve during gradient descent as

$$\frac{d}{dt}\Big(\mathbf{y} - f_{\theta(t)}(X)\Big) = -K^{\mathsf{NTK}}(t) \cdot \Big(\mathbf{y} - f_{\theta(t)}(X)\Big)$$

- For $d_1, \ldots, d_L \to \infty$ and fixed n, K^{NTK} is constant over training.
- Then (diagonalizing K^{NTK}) $\mathbf{y} - f_{\theta(t)}(X) \to 0$ at a different exponential rate along each eigenvector of K^{NTK}.

[Jacot, Gabriel, Hongler '19], [Du et al '19], [Allen-Zhu et al '19], [Lee et al '19], ...

Infinitely wide neural nets are equivalent to kernel linear regression.

Two kernel matrices

2. The **Neural Tangent Kernel**

$$K^{\mathrm{NTK}} = (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n \times n}$$

Training errors evolve during gradient descent as

$$\frac{d}{dt}\Big(\mathbf{y} - f_{\theta(t)}(X)\Big) = -K^{\mathrm{NTK}}(t) \cdot \Big(\mathbf{y} - f_{\theta(t)}(X)\Big)$$

- For $d_1, \ldots, d_L \to \infty$ and fixed n, K^{NTK} is constant over training.
- Then (diagonalizing K^{NTK}) $\mathbf{y} - f_{\theta(t)}(X) \to 0$ at a different exponential rate along each eigenvector of K^{NTK}.

[Jacot, Gabriel, Hongler '19], [Du et al '19], [Allen-Zhu et al '19], [Lee et al '19], ...

Infinitely wide neural nets are equivalent to kernel linear regression. Neural nets of practical width often generalize better than these equivalent kernel models. [Chizat et al '18], [Arora et al '19]

Eigenvalues in the linear width regime

We study the eigenvalue distributions of K^{CK} and K^{NTK}

- In a *linear width* regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ

Eigenvalues in the linear width regime

We study the eigenvalue distributions of K^{CK} and K^{NTK}

- In a *linear width* regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ
- At random (i.i.d. Gaussian) initialization of the weights θ

Eigenvalues in the linear width regime

We study the eigenvalue distributions of K^{CK} and K^{NTK}

- In a *linear width* regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ
- At random (i.i.d. Gaussian) initialization of the weights θ
- Assuming that the training samples $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are approximately pairwise orthogonal, and $\lim \operatorname{spec} X^\top X = \mu_0$

(I'll use "lim spec" to denote weak convergence of the e.s.d.)

Eigenvalues in the linear width regime

We study the eigenvalue distributions of K^{CK} and K^{NTK}

- In a *linear width* regime where $n/d_\ell \to \gamma_\ell \in (0, \infty)$ for each ℓ
- At random (i.i.d. Gaussian) initialization of the weights θ
- Assuming that the training samples $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are approximately pairwise orthogonal, and $\lim \operatorname{spec} X^\top X = \mu_0$

(I'll use "lim spec" to denote weak convergence of the e.s.d.)

Theorem (F., Wang)

For fixed L, almost surely as $n, d_1, \ldots, d_L \to \infty$,

$$\lim \operatorname{spec} K^{CK} = \mu_{CK}, \qquad \lim \operatorname{spec} K^{NTK} = \mu_{NTK}$$

for two probability distributions μ_{CK} and μ_{NTK}. These are defined by μ_0 and properties of $\sigma(x)$.

Approximate pairwise orthogonality

Normalizing training samples such that $\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_n\|^2 \approx 1$, we require

$$|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq \varepsilon_n$$

for each pair $\alpha \neq \beta \in \{1, \ldots, n\}$, where $\varepsilon_n \ll n^{-1/4}$.

Approximate pairwise orthogonality

Normalizing training samples such that $\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_n\|^2 \approx 1$, we require

$$|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq \varepsilon_n$$

for each pair $\alpha \neq \beta \in \{1, \ldots, n\}$, where $\varepsilon_n \ll n^{-1/4}$.

This holds with $\varepsilon_n \approx 1/\sqrt{n}$ if $d_0 \asymp n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are mean-zero independent samples with some concentration. For example:

- $\mathbf{x}_\alpha = \mathbf{z}_\alpha$ where $\mathbf{z}_\alpha$ has i.i.d. subgaussian entries

Approximate pairwise orthogonality

Normalizing training samples such that $\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_n\|^2 \approx 1$, we require

$$|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq \varepsilon_n$$

for each pair $\alpha \neq \beta \in \{1, \ldots, n\}$, where $\varepsilon_n \ll n^{-1/4}$.

This holds with $\varepsilon_n \approx 1/\sqrt{n}$ if $d_0 \asymp n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are mean-zero independent samples with some concentration. For example:

- $\mathbf{x}_\alpha = \mathbf{z}_\alpha$ where $\mathbf{z}_\alpha$ has i.i.d. subgaussian entries
- $\mathbf{x}_\alpha = \Sigma^{1/2}\mathbf{z}_\alpha$ where $\|\Sigma\|$ is bounded

Approximate pairwise orthogonality

Normalizing training samples such that $\|\mathbf{x}_1\|^2, \ldots, \|\mathbf{x}_n\|^2 \approx 1$, we require

$$|\mathbf{x}_\alpha^\top \mathbf{x}_\beta| \leq \varepsilon_n$$

for each pair $\alpha \neq \beta \in \{1, \ldots, n\}$, where $\varepsilon_n \ll n^{-1/4}$.

This holds with $\varepsilon_n \approx 1/\sqrt{n}$ if $d_0 \asymp n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are mean-zero independent samples with some concentration. For example:

- $\mathbf{x}_\alpha = \mathbf{z}_\alpha$ where $\mathbf{z}_\alpha$ has i.i.d. subgaussian entries
- $\mathbf{x}_\alpha = \Sigma^{1/2}\mathbf{z}_\alpha$ where $\|\Sigma\|$ is bounded
- $\mathbf{x}_\alpha = f(\mathbf{z}_\alpha)$ where entries of $\mathbf{z}_\alpha$ satisfy a log-Sobolev inequality, and f is any Lipschitz function

Limit spectral distribution of the CK

Let

$$\mu \mapsto \rho_\gamma^{\mathsf{MP}} \boxtimes \mu$$

be the Marcenko-Pastur map for the spectra of sample covariance matrices with aspect ratio γ.

---

[1]

Limit spectral distribution of the CK

Let

$$\mu \mapsto \rho_\gamma^{\mathsf{MP}} \boxtimes \mu$$

be the Marcenko-Pastur map for the spectra of sample covariance matrices with aspect ratio γ. For $\ell = 1, \ldots, L$, define

$$\mu_\ell = \rho_{\gamma_\ell}^{\mathsf{MP}} \boxtimes \Big((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \Big)$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[1]

[1] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.

Limit spectral distribution of the CK

Let

$$\mu \mapsto \rho_\gamma^{\mathsf{MP}} \boxtimes \mu$$

be the Marcenko-Pastur map for the spectra of sample covariance matrices with aspect ratio γ. For $\ell = 1, \ldots, L$, define

$$\mu_\ell = \rho_{\gamma_\ell}^{\mathsf{MP}} \boxtimes \Big((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \Big)$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[1]

Theorem (F., Wang)

For each $\ell = 1, \ldots, L$, $\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell$. *So* $\lim \operatorname{spec} K^{CK} = \mu_L$.

[1]We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.

Limit spectral distribution of the CK

Let

$$\mu \mapsto \rho_\gamma^{\mathsf{MP}} \boxtimes \mu$$

be the Marcenko-Pastur map for the spectra of sample covariance matrices with aspect ratio γ. For $\ell = 1, \ldots, L$, define

$$\mu_\ell = \rho_{\gamma_\ell}^{\mathsf{MP}} \boxtimes \Big((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\ell-1} \Big)$$

where $b_\sigma = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[\sigma'(\xi)]$.[1]

Theorem (F., Wang)

For each $\ell = 1, \ldots, L$, $\lim \operatorname{spec} X_\ell^\top X_\ell = \mu_\ell$. So $\lim \operatorname{spec} K^{CK} = \mu_L$.

- For one layer, this is closely related to existing results of [Pennington, Worah '17], [Louart, Liao, Couillet '18].
- When $b_\sigma = 0$, each $\mu_\ell = \rho_{\gamma_\ell}^{\mathsf{MP}}$ is a Marcenko-Pastur law. This case was shown (for X with i.i.d. entries) by [Benigni, Péché '19].

---

[1] We normalize σ so that $\mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.

Limit spectral distribution of the NTK

Lemma

There are constants $q_{-1}, \ldots, q_L$ defined by $\sigma(x)$, such that

$$\lim \operatorname{spec} K^{NTK} = \lim \operatorname{spec} \left(q_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} q_\ell X_\ell^\top X_\ell \right)$$

Limit spectral distribution of the NTK

Lemma

There are constants $q_{-1}, \ldots, q_L$ defined by $\sigma(x)$, such that

$$\lim \operatorname{spec} K^{NTK} = \lim \operatorname{spec} \left(q_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} q_\ell X_\ell^\top X_\ell \right)$$

Theorem (F., Wang)

Consider any $\mathbf{z} = (z_{-1}, \ldots, z_L), \mathbf{w} = (w_{-1}, \ldots, w_L)$. Then

$$\frac{1}{n} \operatorname{Tr} \left(z_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} z_\ell X_\ell^\top X_\ell \right)^{-1} \left(w_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} w_\ell X_\ell^\top X_\ell \right)$$

has a deterministic limit $t_L(\mathbf{z}, \mathbf{w})$. A fixed-point equation defines each function t_ℓ in terms of $t_{\ell-1}$.

Limit spectral distribution of the NTK

Lemma

There are constants $q_{-1}, \ldots, q_L$ defined by $\sigma(x)$, such that

$$\lim \operatorname{spec} K^{NTK} = \lim \operatorname{spec} \left(q_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} q_\ell X_\ell^\top X_\ell \right)$$

Theorem (F., Wang)

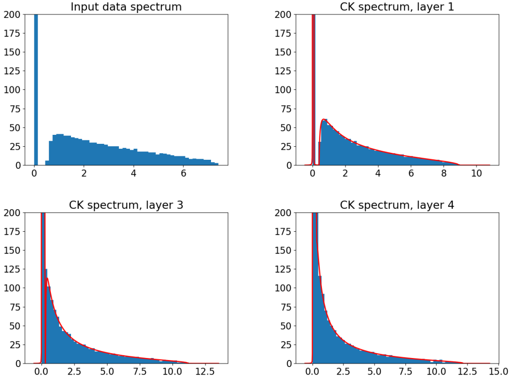
Consider any $\mathbf{z} = (z_{-1}, \ldots, z_L), \mathbf{w} = (w_{-1}, \ldots, w_L)$. Then

$$\frac{1}{n} \operatorname{Tr} \left(z_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} z_\ell X_\ell^\top X_\ell \right)^{-1} \left(w_{-1} \mathrm{Id} + \sum_{\ell=0}^{L} w_\ell X_\ell^\top X_\ell \right)$$

has a deterministic limit $t_L(\mathbf{z}, \mathbf{w})$. A fixed-point equation defines each function t_ℓ in terms of $t_{\ell-1}$. The limit Stieltjes transform for K^{NTK} is then

$$m(z) = t_L\Big((-z + q_{-1}, q_0, \ldots, q_L), (1, 0, \ldots, 0) \Big).$$

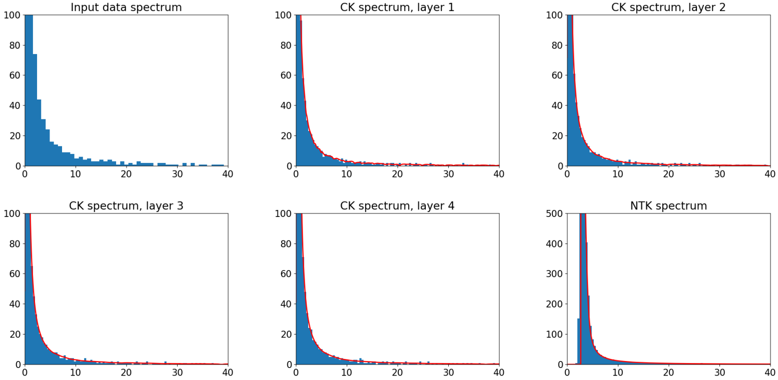
Simulations for i.i.d. Gaussian X



Simulated eigenvalues in blue, limit spectral distribution in red

$\sigma(x) \propto \tan^{-1}(x)$, $L = 5$, $n = 3000$, $d_0 = 1000$, $d_1 = \ldots = d_5 = 6000$

Simulations for input images from CIFAR-10



5000 random training images from CIFAR-10, w/ top 10 PCs removed to improve pairwise orthogonality

$\sigma(x) \propto \tan^{-1}(x)$, $L = 5$, $n = 5000$, $d_0 = 3072$, $d_1 = \ldots = d_5 = 10000$

Main ideas of the analysis

Lemma

Suppose the input data X is ε_n-orthogonal. Then each $X_1, \ldots, X_L$ is $C\varepsilon_n$-orthogonal for a constant $C \equiv C(L) > 0$, w.h.p.

This allows us to induct on the layer ℓ, and analyze each matrix $X_\ell^\top X_\ell$ conditional on $X_0, \ldots, X_{\ell-1}$.

Main ideas of the analysis

Recall $X_\ell = \sigma(W_\ell X_{\ell-1})$, and observe that

- X_ℓ has i.i.d. rows with law $\sigma(\mathbf{w}^\top X_{\ell-1})$, conditional on $X_{\ell-1}$

Main ideas of the analysis

Recall $X_\ell = \sigma(W_\ell X_{\ell-1})$, and observe that

- X_ℓ has i.i.d. rows with law $\sigma(\mathbf{w}^\top X_{\ell-1})$, conditional on $X_{\ell-1}$
- Consequently, $\lim \operatorname{spec} X_\ell^\top X_\ell$ is the Marcenko-Pastur map of

$$\Phi_\ell = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top X_{\ell-1}) \otimes \sigma(\mathbf{w}^\top X_{\ell-1})]$$

[Louart, Liao, Couillet '18]

Main ideas of the analysis

Recall $X_\ell = \sigma(W_\ell X_{\ell-1})$, and observe that

- X_ℓ has i.i.d. rows with law $\sigma(\mathbf{w}^\top X_{\ell-1})$, conditional on $X_{\ell-1}$
- Consequently, $\lim \operatorname{spec} X_\ell^\top X_\ell$ is the Marcenko-Pastur map of

$$\Phi_\ell = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^\top X_{\ell-1}) \otimes \sigma(\mathbf{w}^\top X_{\ell-1})]$$

[Louart, Liao, Couillet '18]

When $X_{\ell-1}$ is ε_n-orthogonal, we show that

$$\frac{1}{n}\left\|\Phi_\ell - \left((1-b_\sigma^2)\mathrm{Id} + b_\sigma^2 X_{\ell-1}^\top X_{\ell-1}\right)\right\|_F^2 \lesssim n \cdot \varepsilon_n^4 \to 0.$$

So $\lim \operatorname{spec} \Phi_\ell = (1-b_\sigma^2) + b_\sigma^2 \mu_{\ell-1}$.

Main ideas of the analysis

To analyze K^{NTK}, we characterize inductively the limit $t_\ell(\mathbf{z}, \mathbf{w})$ of

$$\frac{1}{n} \operatorname{Tr} \left(z_{-1} \mathrm{Id} + \sum_{k=0}^{\ell} z_k X_k^\top X_k \right)^{-1} \left(w_{-1} \mathrm{Id} + \sum_{k=0}^{\ell} w_k X_k^\top X_k \right)$$

Main ideas of the analysis

To analyze K^{NTK}, we characterize inductively the limit $t_\ell(\mathbf{z}, \mathbf{w})$ of

$$\frac{1}{n}\operatorname{Tr}\left(z_{-1}\mathrm{Id} + \sum_{k=0}^{\ell} z_k X_k^\top X_k\right)^{-1}\left(w_{-1}\mathrm{Id} + \sum_{k=0}^{\ell} w_k X_k^\top X_k\right)$$

Remove $X_\ell^\top X_\ell$ from the numerator, by writing this as

$$\frac{w_\ell}{z_\ell} + \frac{1}{n}\operatorname{Tr}\left(A + z_\ell\, X_\ell^\top X_\ell\right)^{-1} M$$

where A, M are linear combinations of $X_0^\top X_0, \ldots, X_{\ell-1}^\top X_{\ell-1}, \mathrm{Id}$.

Main ideas of the analysis

To analyze K^{NTK}, we characterize inductively the limit $t_\ell(\mathbf{z}, \mathbf{w})$ of

$$\frac{1}{n}\operatorname{Tr}\left(z_{-1}\mathrm{Id} + \sum_{k=0}^{\ell} z_k X_k^\top X_k\right)^{-1}\left(w_{-1}\mathrm{Id} + \sum_{k=0}^{\ell} w_k X_k^\top X_k\right)$$

Remove $X_\ell^\top X_\ell$ from the numerator, by writing this as

$$\frac{w_\ell}{z_\ell} + \frac{1}{n}\operatorname{Tr}\left(A + z_\ell X_\ell^\top X_\ell\right)^{-1} M$$

where A, M are linear combinations of $X_0^\top X_0, \ldots, X_{\ell-1}^\top X_{\ell-1}, \mathrm{Id}$.

Conditional on $X_0, \ldots, X_{\ell-1}$, these matrices A and M are deterministic, and X_ℓ is random with i.i.d. rows having second-moment matrix Φ_ℓ.

Main ideas of the analysis

We show an approximation

$$\frac{1}{n}\operatorname{Tr}\left(A + z_\ell X_\ell^\top X_\ell\right)^{-1} M \approx \frac{1}{n}\operatorname{Tr}\left(A + s_\ell^{-1}\Phi_\ell\right)^{-1} M$$

where s_ℓ approximately satisfies the fixed-point equation

$$s_\ell \approx \frac{1}{z_\ell} + \frac{\gamma_\ell}{n}\operatorname{Tr}\left(A + s_\ell^{-1}\Phi_\ell\right)^{-1}\Phi_\ell.$$

This equation depends on the joint spectral limit of (A, Φ_ℓ).

Main ideas of the analysis

We show an approximation

$$\frac{1}{n} \operatorname{Tr} \left(A + z_\ell X_\ell^\top X_\ell \right)^{-1} M \approx \frac{1}{n} \operatorname{Tr} \left(A + s_\ell^{-1} \Phi_\ell \right)^{-1} M$$

where s_ℓ approximately satisfies the fixed-point equation

$$s_\ell \approx \frac{1}{z_\ell} + \frac{\gamma_\ell}{n} \operatorname{Tr} \left(A + s_\ell^{-1} \Phi_\ell \right)^{-1} \Phi_\ell.$$

This equation depends on the joint spectral limit of (A, Φ_ℓ). Applying $\Phi_\ell \approx (1 - b_\sigma^2)\mathrm{Id} + b_\sigma^2 X_{\ell-1}^\top X_{\ell-1}$ and the induction hypothesis for $\ell - 1$, this has a limit in terms of $t_{\ell-1}(\mathbf{z}, \mathbf{w})$.

Main ideas of the analysis

We show an approximation

$$\frac{1}{n}\operatorname{Tr}\left(A + z_\ell X_\ell^\top X_\ell\right)^{-1} M \approx \frac{1}{n}\operatorname{Tr}\left(A + s_\ell^{-1}\Phi_\ell\right)^{-1} M$$

where s_ℓ approximately satisfies the fixed-point equation

$$s_\ell \approx \frac{1}{z_\ell} + \frac{\gamma_\ell}{n}\operatorname{Tr}\left(A + s_\ell^{-1}\Phi_\ell\right)^{-1}\Phi_\ell.$$

This equation depends on the joint spectral limit of (A, Φ_ℓ). Applying $\Phi_\ell \approx (1 - b_\sigma^2)\mathrm{Id} + b_\sigma^2 X_{\ell-1}^\top X_{\ell-1}$ and the induction hypothesis for $\ell - 1$, this has a limit in terms of $t_{\ell-1}(\mathbf{z}, \mathbf{w})$.

We show inductively that the limit equation has a unique fixed point $s_\ell \in \mathbb{C}^+$. This then defines t_ℓ recursively in terms of $t_{\ell-1}$, by

$$t_\ell(\mathbf{z}, \mathbf{w}) = \lim_{n\to\infty}\frac{1}{n}\operatorname{Tr}\left(A + s_\ell^{-1}\Phi_\ell\right)^{-1} M$$

Propagation of "signal" at random initialization

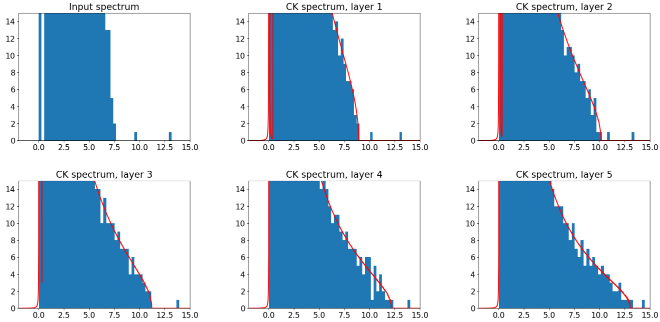
Consider a spiked input matrix

$$X = s_1 \mathbf{u}_1 \mathbf{v}_1^\top + s_2 \mathbf{u}_2 \mathbf{v}_2^\top + \text{i.i.d. Gaussian noise}$$

Propagation of "signal" at random initialization

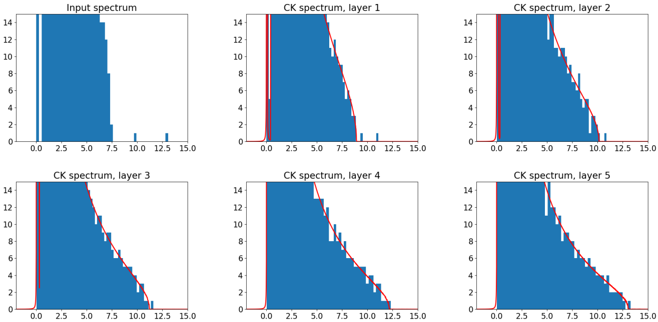
Consider a spiked input matrix

$$X = s_1 \mathbf{u}_1 \mathbf{v}_1^\top + s_2 \mathbf{u}_2 \mathbf{v}_2^\top + \text{i.i.d. Gaussian noise}$$



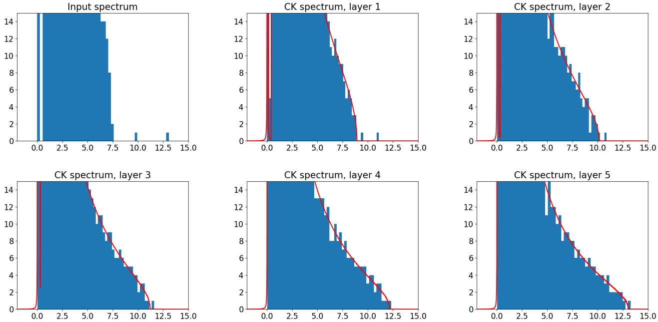
Eigenvalues of $X_\ell^\top X_\ell$, for $\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2$ uniform on the sphere

Propagation of “signal” at random initialization



Eigenvalues of $X_\ell^\top X_\ell$, when $\mathbf{v}_1, \mathbf{v}_2$ are each supported on 20 samples

Propagation of "signal" at random initialization

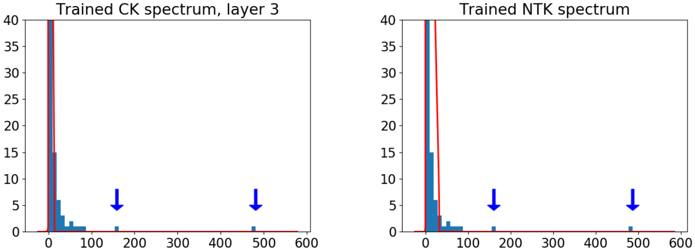


Eigenvalues of $X_\ell^\top X_\ell$, when $\mathbf{v}_1, \mathbf{v}_2$ are each supported on 20 samples

Question: Can we understand the propagation of outlier eigenvalues and eigenvectors through these layers?

Related analysis of Gaussian mixture models for one hidden layer, and other kernels: [Couillet, Benaych-Georges '16], [Liao, Couillet '18]

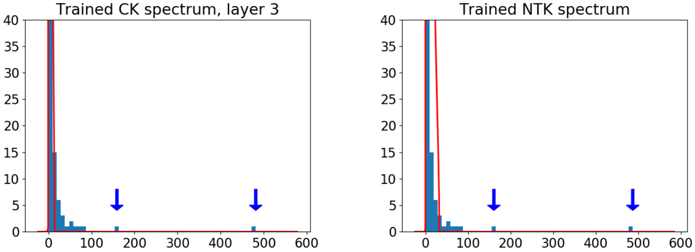
Evolution of spectra over training



Eigenvalues of K^{CK} and K^{NTK} for a trained 3-layer network

$L = 3,\ n = 1000,\ d_0 = 800,\ d_1 = d_2 = d_3 = 800$

Evolution of spectra over training



Eigenvalues of K^{CK} and K^{NTK} for a trained 3-layer network

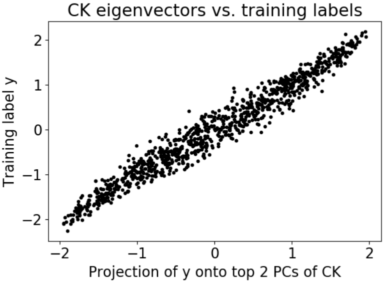
$L = 3,\ n = 1000,\ d_0 = 800,\ d_1 = d_2 = d_3 = 800$

Trained on $(\mathbf{x}_\alpha, y_\alpha)$ pairs where $\mathbf{x}_\alpha$ are uniform on the sphere, and

$$y_\alpha = \sigma(\mathbf{v}^\top \mathbf{x}_\alpha)$$

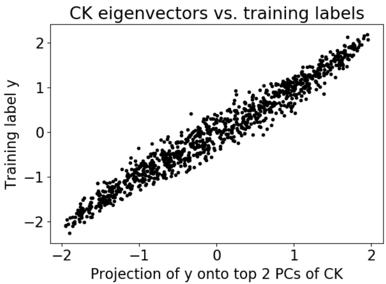
Final prediction-R^2 of the trained model was 0.81. The spectral bulks elongate, and large outliers emerge over training.

Outliers contain information about training labels



Projection of training labels **y** onto top 2 PC's of the trained K^{CK} explains 96% of the variance. The emergence of these outliers is the main mechanism of training in this example.

Outliers contain information about training labels



Projection of training labels **y** onto top 2 PC's of the trained K^{CK} explains 96% of the variance. The emergence of these outliers is the main mechanism of training in this example.

Question: Can we understand the evolutions of K^{CK} and/or K^{NTK} over training, from a spectral perspective?

Related work on the evolution of the NTK in an entrywise size:
[Huang, Yau '19], [Dyer, Gur-Ari '19]

References

Graph matching:

Zhou Fan, Cheng Mao, Yihong Wu, Jiaming Xu, "Spectral graph matching and regularized quadratic relaxations I: The Gaussian model", arxiv:1907.08880.

Zhou Fan, Cheng Mao, Yihong Wu, Jiaming Xu, "Spectral graph matching and regularized quadratic relaxations II: Erdős-Rényi graphs and universality", arxiv:1907.08883.

Neural network kernel matrices:

Zhou Fan, Zhichao Wang, "Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks", arxiv to appear.