Recent Developments in Nonlinear Random Matrices

Yizhe Zhu

Department of Mathematics
UC Irvine

ICERM Random Matrices and Applications
May 20-22, 2024

Nonlinear matrix models

Nonlinear random matrix: an entry-wise nonlinear function applied to a given random matrix

Nonlinear matrix models

Nonlinear random matrix: an entry-wise nonlinear function applied to a given random matrix

- Kernel matrix $\boldsymbol{K}$, $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. e.g., $\boldsymbol{K}_{ij} = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$, $f(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)$. Kernel PCA, kernel SVM, kernel regression.
- Kernel matrices from neural networks.
- Random graphs from nonlinear random matrices.

Random inner product matrix, proportional regime

- Random data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, mean zero, variance 1. $d/n \to \gamma$.
- Random inner product kernel matrix

$$\boldsymbol{K}_{ij} = \begin{cases} \frac{1}{\sqrt{d}} f\left(\frac{1}{\sqrt{d}}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right) & i \neq j \\ 0 & i = j. \end{cases}$$

- $\boldsymbol{x}_i$ Gaussian [Cheng-Singer 13], universality [Do-Vu 13].
- $a = \mathbb{E}_{\xi \sim N(0,1)}[\xi f(\xi)], \quad \nu = \mathbb{E}[f(\xi)^2], \quad \mathbb{E}[f(\xi)] = 0.$
- Limiting spectral distribution of $\boldsymbol{K}$:

$$a(\mu_{\mathrm{MP},\gamma} - 1) \boxplus \sqrt{\gamma^{-1}(\nu - a^2)}\mu_{sc}.$$

- $f(x) = x$, $\nu = a^2$, Marchenko-Pastur law.
- $a = 0$, semicircle.

Concentration

[Fan-Montanari 19] Hermite expansion of $f(x)$:

$$f(x) = \sum_{i=1}^{\infty} a_i h_i(x),$$

$h_i(x)$ normalized hermite polynomials. Decomposition

$$\boldsymbol{K} = \sum_{i=1}^{\infty} a_i \boldsymbol{K}_i.$$

$a_1 \boldsymbol{K}_1$ has a Marchenko-Pastur law, $\sum_{i\geq 2}^{\infty} a_i \boldsymbol{K}_i$ has a semicircle law.

- $\boldsymbol{x}_i$ Gaussian, f is odd, $f(x) = -f(x)$. $\|\boldsymbol{K}\|$ converges to the edge of the limiting spectrum.
- Non-asymptotic bound on $\|\boldsymbol{K}\|$.
- General distribution $\boldsymbol{x}_i$, possible outliers depending on $\mathbb{E}[\mathbf{x}_{ij}^4]$ and a_2.

A different scaling, proportional regime

$$\boldsymbol{K}_{ij} = f\left(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right).$$

A different scaling, proportional regime

$$\boldsymbol{K}_{ij} = f\left(\frac{1}{d}\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right).$$

Theorem (Operator norm approximation, El Karoui 10)

Let $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$*. Assume* $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ *has independent entries with zero mean and unit variance, where* $\mathbb{E}\left|z_i\right|^{4+\eta} \le M$, $\|\boldsymbol{\Sigma}\| \le M$, $\frac{\operatorname{Tr}\boldsymbol{\Sigma}}{d} \to \tau$*. With high probability, when* $n, d \to \infty$ *proportionally,*

$$\left\| \boldsymbol{K} - c_0 \mathbf{1}\mathbf{1}^\top - c_1 \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} - c_2 \mathbf{I}_n \right\| = o(1), \quad \textit{where}$$

$$c_0 = f(0) + \frac{f''(0)}{2}\frac{\operatorname{Tr}(\boldsymbol{\Sigma}^2)}{d^2}, \quad c_1 = f'(0), \quad c_2 = f\left(\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{d}\right) - f(0) - f'(0)\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{d}.$$

$\Longrightarrow$ Marchenko-Pastur law for $\boldsymbol{K}$.

[El Karoui 10]: Taylor expansion

$$\boldsymbol{K}_{ij} = f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right).$$

When $n \asymp d$,

- **Off-diagonal**: $\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d} \approx 0$. Taylor expansion at 0,

$$f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right) \approx f(0) + f'(0)\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d} + \frac{f''(0)}{2}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right)^2 + \frac{f'''(\zeta_{ij})}{6}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right)^3.$$

- **Diagonal**: $\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d} \approx \frac{\operatorname{tr} \boldsymbol{\Sigma}}{d} \approx \tau$. Taylor expansion at τ,

$$f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d}\right) \approx f(\tau) + f'(\zeta_{ii})\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d} - \tau\right).$$

- **Control error terms**: $\|\boldsymbol{K} - c_0 \mathbf{1}\mathbf{1}^\top + c_1 \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + c_2 \mathbf{I}_n\| = o(1)$.

[El Karoui 10]: Taylor expansion

$$\boldsymbol{K}_{ij} = f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right).$$

When $n \asymp d$,

- **Off-diagonal**: $\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d} \approx 0$. Taylor expansion at 0,

$$f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right) \approx f(0) + f'(0)\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d} + \frac{f''(0)}{2}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right)^2 + \frac{f'''(\zeta_{ij})}{6}\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle}{d}\right)^3.$$

- **Diagonal**: $\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d} \approx \frac{\operatorname{tr} \boldsymbol{\Sigma}}{d} \approx \tau$. Taylor expansion at τ,

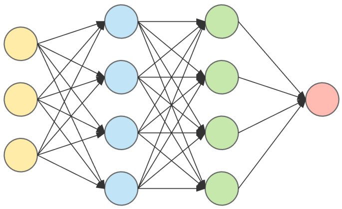
$$f\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d}\right) \approx f(\tau) + f'(\zeta_{ii})\left(\frac{\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}{d} - \tau\right).$$

- **Control error terms**: $\|\boldsymbol{K} - c_0 \mathbf{1}\mathbf{1}^\top + c_1 \frac{\boldsymbol{X}\boldsymbol{X}^\top}{d} + c_2 \mathbf{I}_n\| = o(1)$.
- $\boldsymbol{K}$ has a "low-rank+ bulk + regularizer" structure.

Fully-connected neural network

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

$$f_\theta(\mathbf{x}) = \mathbf{w}^\top \frac{1}{\sqrt{d_L}} \sigma\left(\boldsymbol{W}_L \frac{1}{\sqrt{d_{L-1}}} \sigma\left(\dots \frac{1}{\sqrt{d_2}} \sigma\left(\boldsymbol{W}_2 \frac{1}{\sqrt{d_1}} \sigma(\boldsymbol{W}_1 \mathbf{x}) \right) \right) \right).$$



- $\boldsymbol{W}_1 \in \mathbb{R}^{d_1 \times d_0}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, $\dots$, $\boldsymbol{W}_L \in \mathbb{R}^{d_L \times d_{L-1}}$, and $\mathbf{w} \in \mathbb{R}^{d_L}$. Training parameters: $\theta = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_L, \mathbf{w})$.
- Training samples in a matrix: $\boldsymbol{X}_0 = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{d_0 \times n}$.
- σ: activation function, e.g. $\frac{e^{\alpha x}}{1+e^{\alpha x}}$ (Sigmoid), $|x|$, $\max(0, x)$ (ReLU).
- $\boldsymbol{X}_\ell = \frac{1}{\sqrt{d_\ell}} \sigma\left(\boldsymbol{W}_\ell \boldsymbol{X}_{\ell-1} \right) \in \mathbb{R}^{d_\ell \times n}$, for $1 \le \ell \le L$.

Conjugate kernel

$$\boldsymbol{K}_\ell^{\mathrm{CK}} = \boldsymbol{X}_\ell^\top \boldsymbol{X}_\ell \in \mathbb{R}^{n \times n}$$

- $\boldsymbol{K}_\ell^{\mathrm{CK}}$ governs the properties of random feature regression or network with only the output layer trained.
- At random initialization, its limiting spectrum was studied when all d_i/d_{i-1} and d_0/n are proportional. [Pennington-Worah 17], [Benigni-Péché 19], [Louart-Liao-Couillet 18], [Fan-Wang 20]

Spectrum of conjugate kernels: deterministic data

$$L = 1, \quad \boldsymbol{Y} = \frac{1}{N}\sigma(\boldsymbol{WX})^\top \sigma(\boldsymbol{WX}),$$

$\boldsymbol{W} \in \mathbb{R}^{N\times d}, \boldsymbol{X} \in \mathbb{R}^{d\times n}$. $N/d \to \alpha_1, n/d \to \alpha_2$, $N/n \to \gamma$.

- Deterministic data $\boldsymbol{X}$, $\boldsymbol{W}$ Gaussian, Lipschitz activation σ. Row vectors of $\sigma(\boldsymbol{WX})$ are independent. [Louart-Liao-Couillet 18]
- Limiting spectral distribution of $\boldsymbol{Y}$ is $\mu_{MP} \boxtimes \mu_{\boldsymbol{\Phi}}$, where $\mu_{\boldsymbol{\Phi}}$ is the limiting spectral distribution of

$$\boldsymbol{\Phi} = \mathbb{E}_{\boldsymbol{W}}[\boldsymbol{Y}] = \mathbb{E}_{\boldsymbol{w}}[\sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top \sigma(\boldsymbol{w}^\top \boldsymbol{X})] \in \mathbb{R}^{n\times n}.$$

- Same limiting ESD as a linear model $\frac{1}{N}\boldsymbol{P}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{P}$ with $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{\Phi}$, $\boldsymbol{P} \in \mathbb{R}^{d\times n}$.
- Key step: concentration of random quadratic forms $\sigma(\boldsymbol{w}^\top \boldsymbol{X})\boldsymbol{A}\, \sigma(\boldsymbol{w}^\top \boldsymbol{X})^\top$ for deterministic $\boldsymbol{A}$.

Conjugate kernel, approximately orthonormal data

- When columns of $\boldsymbol{X}$ are *approximately orthonormal* [Fang-Wang 20],

$$\mu = \rho_\gamma^{\mathsf{MP}} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\boldsymbol{X}^\top \boldsymbol{X}} \right),$$

  where $b_\sigma = \mathbb{E}_{\xi \sim \mathbb{N}(0,1)}[\sigma'(\xi)], \quad \mathbb{E}[\sigma(\xi)] = 0, \; \mathbb{E}[\sigma(\xi)^2] = 1.$

- From [LLC 18] and a first order approximation $\boldsymbol{\Phi} \approx (1 - b_\sigma^2)\mathbf{I} + b_\sigma^2 \boldsymbol{X}\boldsymbol{X}^\top$.

Conjugate kernel, approximately orthonormal data

- When columns of $\boldsymbol{X}$ are *approximately orthonormal* [Fang-Wang 20],

$$\mu = \rho_\gamma^{\mathsf{MP}} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\boldsymbol{X}^\top \boldsymbol{X}} \right),$$

  where $b_\sigma = \mathbb{E}_{\xi \sim \mathbb{N}(0,1)}[\sigma'(\xi)], \quad \mathbb{E}[\sigma(\xi)] = 0,\ \mathbb{E}[\sigma(\xi)^2] = 1.$

- From [LLC 18] and a first order approximation $\boldsymbol{\Phi} \approx (1 - b_\sigma^2)\mathbf{I} + b_\sigma^2 \boldsymbol{X}\boldsymbol{X}^\top$.

Conjugate kernel, approximately orthonormal data

- When columns of $\boldsymbol{X}$ are *approximately orthonormal* [Fang-Wang 20],

$$\mu = \rho_\gamma^{\mathrm{MP}} \boxtimes \left((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\boldsymbol{X}^\top \boldsymbol{X}} \right),$$

  where $b_\sigma = \mathbb{E}_{\xi \sim \mathbb{N}(0,1)}[\sigma'(\xi)], \quad \mathbb{E}[\sigma(\xi)] = 0,\ \mathbb{E}[\sigma(\xi)^2] = 1.$
- From [LLC 18] and a first order approximation $\boldsymbol{\Phi} \approx (1 - b_\sigma^2)\mathbf{I} + b_\sigma^2 \boldsymbol{X}\boldsymbol{X}^\top$.
- Can be extended to L layers. Approximate orthogonality propagates through the nonlinear map $\boldsymbol{X}_{\ell-1} \to \boldsymbol{X}_\ell = \frac{1}{\sqrt{d_\ell}} \sigma(\boldsymbol{W}\boldsymbol{X}_{\ell-1})$.

Conjugate kernel, approximately orthonormal data

- When columns of $\boldsymbol{X}$ are *approximately orthonormal* [Fang-Wang 20],

$$\mu = \rho_\gamma^{\mathsf{MP}} \boxtimes \Big((1 - b_\sigma^2) + b_\sigma^2 \cdot \mu_{\boldsymbol{X}^\top \boldsymbol{X}} \Big),$$

  where $b_\sigma = \mathbb{E}_{\xi \sim \mathbb{N}(0,1)}[\sigma'(\xi)]$, $\quad \mathbb{E}[\sigma(\xi)] = 0$, $\mathbb{E}[\sigma(\xi)^2] = 1$.

- From [LLC 18] and a first order approximation $\boldsymbol{\Phi} \approx (1 - b_\sigma^2)\mathbf{I} + b_\sigma^2 \boldsymbol{X}\boldsymbol{X}^\top$.
- Can be extended to L layers. Approximate orthogonality propagates through the nonlinear map $\boldsymbol{X}_{\ell-1} \to \boldsymbol{X}_\ell = \frac{1}{\sqrt{d_\ell}} \sigma(\boldsymbol{W}\boldsymbol{X}_{\ell-1})$.
- $L = 1$, $N \gg n$, deformed semicircle law for $\boldsymbol{Y}$ [Wang-Z. 24]. Training and generalization error with deterministic data [Wang-Z. 23, Latourelle-Vigeant Paquette 23].

Spectrum of conjugate kernels: random data

- **Universality** [Benigni-Péché 19], i.i.d. entries in $\boldsymbol{X}, \boldsymbol{W}$ with mean zero, variance 1, general distributions, σ analytic.
- The limiting spectral distribution depends on

$$\theta_1(\sigma) = \mathbb{E}_{\xi \sim N(0,1)}[\sigma^2(\xi)], \qquad \theta_2(\sigma) = \left(\mathbb{E}_{\xi \sim N(0,1)}[\sigma'(\xi)]\right)^2.$$

- Same as the limiting ESD of an information-plus-noise matrix:

$$\boldsymbol{M} = \frac{1}{N}\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)^\top,$$

  where $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}$ are Gaussian with i.i.d. entries. (*Gaussian Equivalence*)

Spectrum of conjugate kernels: random data

- **Universality** [Benigni-Péché 19], i.i.d. entries in $\boldsymbol{X}, \boldsymbol{W}$ with mean zero, variance 1, general distributions, σ analytic.
- The limiting spectral distribution depends on

$$\theta_1(\sigma) = \mathbb{E}_{\xi \sim N(0,1)}[\sigma^2(\xi)], \qquad \theta_2(\sigma) = \left(\mathbb{E}_{\xi \sim N(0,1)}[\sigma'(\xi)]\right)^2.$$

- Same as the limiting ESD of an information-plus-noise matrix:

$$\boldsymbol{M} = \frac{1}{N}\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)^\top,$$

  where $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}$ are Gaussian with i.i.d. entries. (*Gaussian Equivalence*)
- Outliers depending on Hermite coefficients similar to [Fan-Montanari 19].
- Extension to L layers: when $\theta_2(\sigma) = 0$, μ_L is Marchenko-Pastur.

Spectrum of conjugate kernels: random data

- **Universality** [Benigni-Péché 19], i.i.d. entries in $\boldsymbol{X}, \boldsymbol{W}$ with mean zero, variance 1, general distributions, σ analytic.
- The limiting spectral distribution depends on

$$\theta_1(\sigma) = \mathbb{E}_{\xi \sim N(0,1)}[\sigma^2(\xi)], \qquad \theta_2(\sigma) = \left(\mathbb{E}_{\xi \sim N(0,1)}[\sigma'(\xi)]\right)^2.$$

- Same as the limiting ESD of an information-plus-noise matrix:

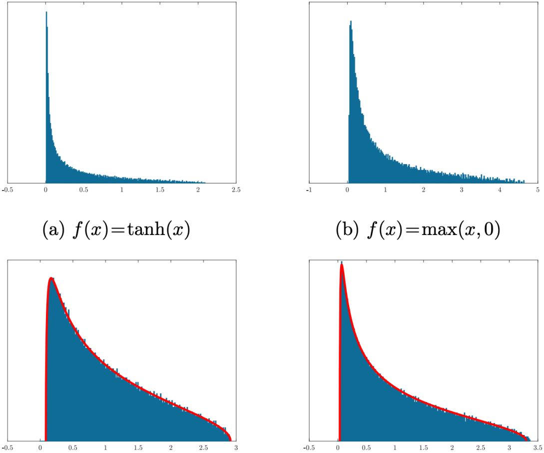
$$\boldsymbol{M} = \frac{1}{N}\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)\left(\frac{\sqrt{\theta_2}}{\sqrt{d}}\boldsymbol{W}\boldsymbol{X} + \sqrt{\theta_1 - \theta_2}\boldsymbol{Z}\right)^\top,$$

where $\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}$ are Gaussian with i.i.d. entries. (*Gaussian Equivalence*)

- Outliers depending on Hermite coefficients similar to [Fan-Montanari 19].
- Extension to L layers: when $\theta_2(\sigma) = 0$, μ_L is Marchenko-Pastur.
- Matched with [Fan-Wang 20] if $\boldsymbol{X}\boldsymbol{X}^\top$ has a Marchenko-Pastur distribution.

Question: a unified proof for two types of conditions on σ and $\boldsymbol{W}$?

Spectrum of conjugate kernels, random data



(a) $f(x) = \tanh(x)$

(b) $f(x) = \max(x, 0)$

(c) $f(x) = \cos(x)$

(d) $f(x) = x^3 - 3x$

Figure: [Benigni-Péché 21]

Neural tangent kernel

$$
\begin{aligned}
K^{\mathrm{NTK}} :=& (\nabla_\theta f_\theta(X))^\top (\nabla_\theta f_\theta(X)) \in \mathbb{R}^{n\times n} \\
=& X_L^\top X_L + \sum_{\ell=1}^{L} (S_\ell^\top S_\ell) \odot (X_{\ell-1}^\top X_{\ell-1})
\end{aligned}
$$

When $L = 1$,

$$
K^{\mathrm{NTK}} = X_1^\top X_1 + X^\top X \odot \left(\frac{1}{d_1} \sigma'(WX)^\top \mathrm{diag}(\mathbf{w})^2 \sigma'(WX) \right).
$$

- Training errors evolved during gradient descent is governed by K^{NTK}. For $d_1 \to \infty$ and fixed n, K^{NTK} converges to its expectation and is fixed over training in the infinite width limit.
- The smallest singular value of K^{NTK} controls the global convergence of gradient descent.

[Jacot, Gabriel, Hongler 18], [Chizat et al 18], [Du et al 19], [Allen-Zhu et al 19], [Lee et al 19], [Arora et al 19], [Oymak-Soltanolkotabi 20], [Adlam et al 20], [Fan, Wang 20], [Montanari Zhong 22], [Bombari-Amani-Mondelli 22] ...

Random feature regression

A two-layer neural network $f : \mathbb{R}^d \to \mathbb{R}$ at random initialization

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{n}} \boldsymbol{\theta}^\top \sigma(\boldsymbol{W}\boldsymbol{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{N} \boldsymbol{\theta}_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}).$$

- $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times d}$: weight matrix with i.i.d. $N(0,1)$ entries.
- Training the output layer weight= linear regression with respect to random features $\phi(\boldsymbol{x}_i) = \sigma(\boldsymbol{W}\boldsymbol{x}_i) \in \mathbb{R}^N$.

[Ghorbani-Mei-Misiakiewicz-Montanari 21, Mei-Montanari 22, Misiakiewicz 22, Hu-Lu 22, Montanari-Zhong 22],...

Random feature ridge regression (RFRR)

Training data $(\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)$, $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$.

- The loss function is defined by

$$L(\boldsymbol{\theta}) = \frac{1}{n}\|f(\boldsymbol{X}) - \boldsymbol{y}\|^2 + \frac{\lambda}{n}\|\boldsymbol{\theta}\|^2.$$

- Then the optimal predictor for RFRR is given by

$$\hat{f}_\lambda^{(\mathrm{RF})}(\boldsymbol{x}) = \boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}_N + \lambda \mathrm{Id})^{-1}\boldsymbol{y},$$

where $\boldsymbol{K}_N$ is the empirical *conjugate kernel matrix*:

$$\boldsymbol{K}_N = \frac{1}{N}\sigma\left(\boldsymbol{W}\boldsymbol{X}\right)^\top \sigma\left(\boldsymbol{W}\boldsymbol{X}\right) \in \mathbb{R}^{n\times n}.$$

Random feature ridge regression (RFRR)

Training data $(\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)$, $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$.

- The loss function is defined by

$$L(\boldsymbol{\theta}) = \frac{1}{n}\|f(\boldsymbol{X}) - \boldsymbol{y}\|^2 + \frac{\lambda}{n}\|\boldsymbol{\theta}\|^2.$$

- Then the optimal predictor for RFRR is given by

$$\hat{f}_\lambda^{(\mathrm{RF})}(\boldsymbol{x}) = \boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}_N + \lambda \mathrm{Id})^{-1}\boldsymbol{y},$$

  where $\boldsymbol{K}_N$ is the empirical *conjugate kernel matrix*:

$$\boldsymbol{K}_N = \frac{1}{N}\sigma(\boldsymbol{W}\boldsymbol{X})^\top \sigma(\boldsymbol{W}\boldsymbol{X}) \in \mathbb{R}^{n \times n}.$$

- *Training error*:

$$E_{\mathrm{train}}^{(RF,\lambda)} := \frac{1}{n}\|\hat{f}_\lambda^{(RF)}(\boldsymbol{X}) - \mathbf{y}\|_2^2 = \frac{\lambda^2}{n}\|(\boldsymbol{K}_N + \lambda \mathrm{Id})^{-1}\mathbf{y}\|^2.$$

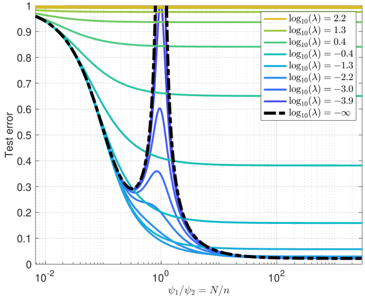
- *Test/ generalization error*: $\boldsymbol{x}$ sampled from the same distribution as training data,

$$\mathcal{R}(\hat{f}) := \mathbb{E}_{\mathbf{x}}[|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})|^2].$$

Double descent for generalization error

[Mei-Montanari 22] (informal):

- Assume $\boldsymbol{w}_i, \boldsymbol{x}_i$ are i.i.d. uniformly distributed on $\mathbb{S}^{d-1}$, $y_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \varepsilon_i$.
- $N/d \to \psi_1, n/d \to \psi_2$, $\lim_n \mathcal{R}(\hat{f})$ is a function of λ, ψ_1, ψ_2 and other model parameters.



Question: Universality for general weights/data distributions?

Kernel ridge regression

- Consider the empirical Risk Minimization (ERM)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}},$$

  where $\lambda \geq 0$ and $\mathcal{H}$ is the Reproducing Kernel Hilbert Space for $k(\cdot, \cdot)$.

- Kernel ridge regression's predictor:

$$\hat{f}_{\lambda}^{(\mathrm{K})}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \lambda \mathrm{Id})^{-1} \boldsymbol{y},$$

  where $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{X}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_1)]$ and $(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}))_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $1 \leq i, j \leq n$ and $\boldsymbol{x}, \boldsymbol{x}_i \in \mathbb{R}^d$.

Kernel ridge regression

- Consider the empirical Risk Minimization (ERM)

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \left\| f \right\|_{\mathcal{H}},$$

  where $\lambda \geq 0$ and $\mathcal{H}$ is the Reproducing Kernel Hilbert Space for $k(\cdot,\cdot)$.

- Kernel ridge regression's predictor:

$$\hat{f}_{\lambda}^{(\mathrm{K})}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda \mathrm{Id})^{-1}\boldsymbol{y},$$

  where $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) = [k(\boldsymbol{x},\boldsymbol{x}_1),\ldots,k(\boldsymbol{x},\boldsymbol{x}_1)]$ and $(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}))_{i,j} = k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for $1 \leq i,j \leq n$ and $\boldsymbol{x},\boldsymbol{x}_i \in \mathbb{R}^d$.

- When $N \gg n$, random feature regression can be approximated by KRR.

Kernel ridge regression

- Consider the empirical Risk Minimization (ERM)

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}},$$

  where $\lambda \geq 0$ and $\mathcal{H}$ is the Reproducing Kernel Hilbert Space for $k(\cdot,\cdot)$.

- Kernel ridge regression's predictor:

$$\hat{f}_{\lambda}^{(\mathrm{K})}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \lambda \mathrm{Id})^{-1}\boldsymbol{y},$$

  where $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{X}) = [k(\boldsymbol{x},\boldsymbol{x}_1),\ldots,k(\boldsymbol{x},\boldsymbol{x}_1)]$ and $(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}))_{i,j} = k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for $1 \leq i,j \leq n$ and $\boldsymbol{x},\boldsymbol{x}_i \in \mathbb{R}^d$.

- When $N \gg n$, random feature regression can be approximated by KRR.

[Bartlett-Montanari-Rakhlin 21]: For $\boldsymbol{K}_{ij} = f(\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle/d)$, $n \asymp d$, KRR is asymptotically equivalent to linear ridge regression with a different ridge parameter.

proved for subgaussian data $\boldsymbol{x}_i$ with general covariance $\boldsymbol{\Sigma}$.

Beyond $n \asymp d$, polynomial regime $n \asymp d^k$

Beyond $n \asymp d$, polynomial regime $n \asymp d^k$

- A simple example: when $f(x) = x^k$, $\boldsymbol{K}(x_i, x_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^k = \langle \boldsymbol{x}_i^{\otimes k}, \boldsymbol{x}_j^{\otimes k} \rangle$.
- Let $\boldsymbol{Y} = [\boldsymbol{x}_1^{\otimes k}, \cdots, \boldsymbol{x}_n^{\otimes k}] \in \mathbb{R}^{d^k \times n}$. Then $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ has a Marchenko-Pastur law when $n \asymp d^k$ [Yaskov 23]. Connection to random tensor models [Bryson-Vershynin-Zhao 21].

Beyond $n \asymp d$, polynomial regime $n \asymp d^k$

- A simple example: when $f(x) = x^k$, $\boldsymbol{K}(x_i, x_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^k = \langle \boldsymbol{x}_i^{\otimes k}, \boldsymbol{x}_j^{\otimes k} \rangle$.
- Let $\boldsymbol{Y} = [\boldsymbol{x}_1^{\otimes k}, \cdots, \boldsymbol{x}_n^{\otimes k}] \in \mathbb{R}^{d^k \times n}$. Then $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ has a Marchenko-Pastur law when $n \asymp d^k$ [Yaskov 23]. Connection to random tensor models [Bryson-Vershynin-Zhao 21].

Universality [Lu-Yau 22, Dubova-Lu-McKenna-Yau 23]

$f(x) = \sum_{k=0}^{L} c_k h_k(x)$. $\boldsymbol{K}_{ij} = \frac{1}{\sqrt{n}} f(\frac{1}{\sqrt{d}} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) \mathbf{1}\{i \neq j\}$. $\frac{n}{d^\ell} \to \kappa > 0$. $\boldsymbol{x}_i$ has i.i.d. entries with all finite moments.

- When ℓ is an integer, the limiting law is the free convolution of the semicirle law and Marchenko-Pastur law.
- When ℓ is not an integer, the limiting law is semicircle.

Beyond $n \asymp d$, polynomial regime $n \asymp d^k$

- A simple example: when $f(x) = x^k$, $\boldsymbol{K}(x_i, x_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^k = \langle \boldsymbol{x}_i^{\otimes k}, \boldsymbol{x}_j^{\otimes k} \rangle$.
- Let $\boldsymbol{Y} = [\boldsymbol{x}_1^{\otimes k}, \cdots, \boldsymbol{x}_n^{\otimes k}] \in \mathbb{R}^{d^k \times n}$. Then $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ has a Marchenko-Pastur law when $n \asymp d^k$ [Yaskov 23]. Connection to random tensor models [Bryson-Vershynin-Zhao 21].

Universality [Lu-Yau 22, Dubova-Lu-McKenna-Yau 23]

$f(x) = \sum_{k=0}^{L} c_k h_k(x)$. $\boldsymbol{K}_{ij} = \frac{1}{\sqrt{n}} f(\frac{1}{\sqrt{d}} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) \mathbf{1}\{i \neq j\}$. $\frac{n}{d^\ell} \to \kappa > 0$. $\boldsymbol{x}_i$ has i.i.d. entries with all finite moments.

- When ℓ is an integer, the limiting law is the free convolution of the semicirle law and Marchenko-Pastur law.
- When ℓ is not an integer, the limiting law is semicircle.

Beyond $n \asymp d$, polynomial regime $n \asymp d^k$

- A simple example: when $f(x) = x^k$, $\boldsymbol{K}(x_i, x_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^k = \langle \boldsymbol{x}_i^{\otimes k}, \boldsymbol{x}_j^{\otimes k} \rangle$.
- Let $\boldsymbol{Y} = [\boldsymbol{x}_1^{\otimes k}, \cdots, \boldsymbol{x}_n^{\otimes k}] \in \mathbb{R}^{d^k \times n}$. Then $\boldsymbol{K} = \boldsymbol{Y}^\top \boldsymbol{Y}$ has a Marchenko-Pastur law when $n \asymp d^k$ [Yaskov 23]. Connection to random tensor models [Bryson-Vershynin-Zhao 21].

Universality [Lu-Yau 22, Dubova-Lu-McKenna-Yau 23]

$f(x) = \sum_{k=0}^{L} c_k h_k(x)$. $\boldsymbol{K}_{ij} = \frac{1}{\sqrt{n}} f(\frac{1}{\sqrt{d}} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) \mathbf{1}\{i \neq j\}$. $\frac{n}{d^\ell} \to \kappa > 0$. $\boldsymbol{x}_i$ has i.i.d. entries with all finite moments.

- When ℓ is an integer, the limiting law is the free convolution of the semicirle law and Marchenko-Pastur law.
- When ℓ is not an integer, the limiting law is semicircle.

heuristics: When $\ell \in \mathbb{Z}$, $\boldsymbol{K} = \sum_{i=1}^{L} a_i \boldsymbol{K}_i$, each $\boldsymbol{K}_i$ approximately independent. $\sum_{i=1}^{\ell-1} \boldsymbol{K}_i$ is low-rank, $\boldsymbol{K}_\ell$ has a Marchenko-Pastur law, $\sum_{i=\ell+1}^{L} a_i \boldsymbol{K}_i$ has a semicircle law.

Kernel regression in the polynomial regime

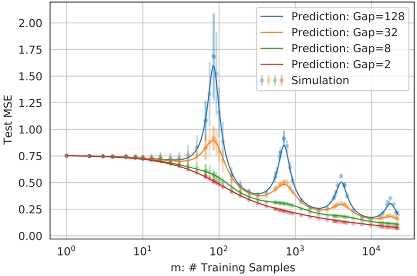
[Xiao-Hu-Misiakiewicz-Lu-Pennington 22]

- When $n \asymp d^{\ell}, \ell \in \mathbb{Z}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- $\boldsymbol{K}_{ij} = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$. Different scaling compared to [LY22, DLMY23]. $\boldsymbol{K}$ has a Marchenko-Pastur law. Generalization of [El Karoui 10].

Kernel regression in the polynomial regime

[Xiao-Hu-Misiakiewicz-Lu-Pennington 22]

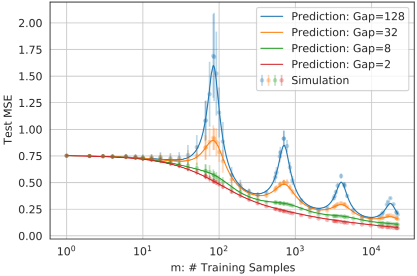
- When $n \asymp d^\ell, \ell \in \mathbb{Z}$, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- $\boldsymbol{K}_{ij} = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$. Different scaling compared to [LY22, DLMY23]. $\boldsymbol{K}$ has a Marchenko-Pastur law. Generalization of [El Karoui 10].
- Multiple decents in the generalization error. (Informally) if $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, KRR with $n \asymp d^\ell$ many samples can learn the first ℓ-th degree components of f_*.



Kernel regression in the polynomial regime

[Xiao-Hu-Misiakiewicz-Lu-Pennington 22]

- When $n \asymp d^\ell, \ell \in \mathbb{Z}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- $\boldsymbol{K}_{ij} = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$. Different scaling compared to [LY22, DLMY23]. $\boldsymbol{K}$ has a Marchenko-Pastur law. Generalization of [El Karoui 10].
- Multiple decents in the generalization error. (Informally) if $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$, KRR with $n \asymp d^\ell$ many samples can learn the first ℓ-th degree components of f_*.



- Random feature regression $N \asymp d^{\kappa_1}, n \asymp d^{\kappa_2}$ [Hu-Lu-Misiakiewicz 24].

Nonlinear spiked model

- Spiked Wigner model $\boldsymbol{Y} = \frac{1}{\sqrt{n}}\boldsymbol{A} + \frac{\gamma}{n}\boldsymbol{x}\boldsymbol{x}^\top$, BBP transition at $|\gamma| = 1$.
  [Baik-Ben Arous-Péché 05, Benaych-Georges, Nadakuditi 11].

Nonlinear spiked model

- Spiked Wigner model $\boldsymbol{Y} = \frac{1}{\sqrt{n}}\boldsymbol{A} + \frac{\gamma}{n}\boldsymbol{x}\boldsymbol{x}^\top$, BBP transition at $|\gamma| = 1$. [Baik-Ben Arous-Péché 05, Benaych-Georges, Nadakuditi 11].
- Nonlinear spiked Wigner model [Guionnet-Ko-Krzakala-Mergny-Zdebrová 23]

$$\boldsymbol{Y} = \frac{1}{\sqrt{n}}\left[f\left(\boldsymbol{Z} + \frac{\gamma(n)}{\sqrt{n}}\boldsymbol{x}\boldsymbol{x}^\top\right) - \mathbb{E}f(\boldsymbol{Z})\right].$$

Nonlinear spiked model

- Spiked Wigner model $\boldsymbol{Y} = \frac{1}{\sqrt{n}}\boldsymbol{A} + \frac{\gamma}{n}\boldsymbol{x}\boldsymbol{x}^\top$, BBP transition at $|\gamma| = 1$. [Baik-Ben Arous-Péché 05, Benaych-Georges, Nadakuditi 11].
- Nonlinear spiked Wigner model [Guionnet-Ko-Krzakala-Mergny-Zdebrová 23]

$$\boldsymbol{Y} = \frac{1}{\sqrt{n}}\left[f\left(\boldsymbol{Z} + \frac{\gamma(n)}{\sqrt{n}}\boldsymbol{x}\boldsymbol{x}^\top\right) - \mathbb{E}f(\boldsymbol{Z})\right].$$

When $\boldsymbol{Z}$ is Gaussian, $\boldsymbol{x}$ is random, phase transition of spikes happens at

$$\gamma(n) \asymp n^{\frac{1}{2}(1-\frac{1}{k_*})},$$

where k_* is the degree of the first nonzero hermite polynomial in the expansion of f. $k_* = 1$ when f is linear.

Nonlinear spiked model

- Spiked Wigner model $\boldsymbol{Y} = \frac{1}{\sqrt{n}}\boldsymbol{A} + \frac{\gamma}{n}\boldsymbol{x}\boldsymbol{x}^\top$, BBP transition at $|\gamma| = 1$. [Baik-Ben Arous-Péché 05, Benaych-Georges, Nadakuditi 11].
- Nonlinear spiked Wigner model [Guionnet-Ko-Krzakala-Mergny-Zdebrová 23]

$$\boldsymbol{Y} = \frac{1}{\sqrt{n}}\left[f\left(\boldsymbol{Z} + \frac{\gamma(n)}{\sqrt{n}}\boldsymbol{x}\boldsymbol{x}^\top\right) - \mathbb{E}f(\boldsymbol{Z})\right].$$

  When $\boldsymbol{Z}$ is Gaussian, $\boldsymbol{x}$ is random, phase transition of spikes happens at

$$\gamma(n) \asymp n^{\frac{1}{2}(1-\frac{1}{k_*})},$$

  where k_* is the degree of the first nonzero hermite polynomial in the expansion of f. $k_* = 1$ when f is linear.
- Nonlinear spiked covariance model and connection to neural networks [Ba-Erdogdu-Suzuki-Wang-Wu 23, Wang-Wu-Fan 24].

Other topics

- Feature learning with gradient descent [Ba-Murat-Erdogdu-Suzuki-Wang-Wu-Yang 22]
- Spectrum of empirical Hessian after SGD [Ben Arous-Gheissari-Huang-Jagannath 23]
- Generalization error of SGD in high dimensions [Paquette-Paquette-Adlam-Pennington 22]
- Gaussian equivalence [Hu-Lu 22], [Goldt-Loureiro-Reeves-Krzakala-Mzard-Zdeborová 21], [Montanari-Ruan-Saeed-Sohn 23],. . .
- Benign overfitting [Bartlett-Long-Lugosi-Tsigler 20], [Tsigler-Bartlett 23], [Koehler-Zhou-Sutherland-Srebro 21],. . .

Random geometric graphs $G(n, d, p)$

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ sampled i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- (i, j) are connected if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau$. Choose $\tau = \tau(p, d)$ to match the edge density of a $G(n, p)$.

Random geometric graphs $G(n, d, p)$

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ sampled i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- (i, j) are connected if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau$. Choose $\tau = \tau(p, d)$ to match the edge density of a $G(n, p)$.
- Adjacency matrix $A_{ij} = \mathbf{1}\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau\}$: random kernel matrix with $f(x) = \mathbf{1}\{x \geq \tau\}$.

Random geometric graphs $G(n, d, p)$

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sampled i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- (i, j) are connected if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau$. Choose $\tau = \tau(p, d)$ to match the edge density of a $G(n, p)$.
- Adjacency matrix $A_{ij} = \mathbf{1}\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau\}$: random kernel matrix with $f(x) = \mathbf{1}\{x \geq \tau\}$.

When do random geometric graphs lose geometry?

Random geometric graphs $G(n, d, p)$

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sampled i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- (i, j) are connected if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau$. Choose $\tau = \tau(p, d)$ to match the edge density of a $G(n, p)$.
- Adjacency matrix $A_{ij} = \mathbf{1}\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau\}$: random kernel matrix with $f(x) = \mathbf{1}\{x \geq \tau\}$.

When do random geometric graphs lose geometry?

Theorem (Bubeck-Ding-Eldan-Rácz 16)

Let $p \in (0, 1)$ be fixed.

- *When $d \gg n^3$, $\mathrm{TV}(G(n, p), G(n, p, d)) \to 0$.*
- *When $d \ll n^3$, $\mathrm{TV}(G(n, p), G(n, p, d)) \to 1$.*

Random geometric graphs $G(n, d, p)$

- $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sampled i.i.d. uniformly on $\mathbb{S}^{d-1}$.
- (i, j) are connected if $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau$. Choose $\tau = \tau(p, d)$ to match the edge density of a $G(n, p)$.
- Adjacency matrix $A_{ij} = \mathbf{1}\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \tau\}$: random kernel matrix with $f(x) = \mathbf{1}\{x \geq \tau\}$.

When do random geometric graphs lose geometry?

Theorem (Bubeck-Ding-Eldan-Rácz 16)

Let $p \in (0, 1)$ be fixed.

- *When $d \gg n^3$, $\mathrm{TV}(G(n,p), G(n,p,d)) \to 0$.*
- *When $d \ll n^3$, $\mathrm{TV}(G(n,p), G(n,p,d)) \to 1$.*

Connected to TV distance between $\mathrm{GOE}(n)$ and $\mathrm{Wishart}(n, d)$ [Jiang-Li 15]. Detect local dependence by counting signed triangles.

Random geometric graphs $G(n, d, p)$

- **Question**: Sparse regime $p = \frac{c}{n}$. $d \asymp \log^3(n)$ is the conjectured threshold in [BDER16].
- $p = \frac{c}{n}$: distinguishable when $d \ll \log^3(n)$ [BDER 16], indistinguishable when $c > 1, d \gg \log^{36}(n)$ [Liu-Mohanty-Schramm-Yang 22].

Random geometric graphs $G(n, d, p)$

- **Question**: Sparse regime $p = \frac{c}{n}$. $d \asymp \log^3(n)$ is the conjectured threshold in [BDER16].
- $p = \frac{c}{n}$: distinguishable when $d \ll \log^3(n)$ [BDER 16], indistinguishable when $c > 1, d \gg \log^{36}(n)$ [Liu-Mohanty-Schramm-Yang 22].
- Spectral gap of $\boldsymbol{A}$ in certain regimes [Liu-Mohanty-Schramm-Yang 22, Bagachev-Bresler 24].

Random geometric graphs $G(n, d, p)$

- **Question**: Sparse regime $p = \frac{c}{n}$. $d \asymp \log^3(n)$ is the conjectured threshold in [BDER16].
- $p = \frac{c}{n}$: distinguishable when $d \ll \log^3(n)$ [BDER 16], indistinguishable when $c > 1, d \gg \log^{36}(n)$ [Liu-Mohanty-Schramm-Yang 22].
- Spectral gap of $\boldsymbol{A}$ in certain regimes [Liu-Mohanty-Schramm-Yang 22, Bagachev-Bresler 24].
- Geometric block model: $\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$ are drawn from a Gaussian mixture [Li-Schramm 24]. Testing geometry, community recovery/clustering.

Conclusions

- In the proportional regime $n \asymp d$, a nonlinear random matrix model behaves like another linear random matrix model.
- A polynomial regime $n \asymp d^{\ell}$ appears for nonlinear models, new phenomena in the spectrum and regression performance.
- **Question**: Universality and structured data for generalization error.
- **Question**: spectrum of geometric random graphs.