Random Matrix Theory for Modern Machine Learning: New Intuitions, Improved Methods, and Beyond: Part 4

Short Course @ Institut de Mathématiques de Toulouse, France

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

July 4th, 2024

Outline

1. Linearization of Nonlinear Models
   - Taylor expansion
   - Orthogonal polynomial

2. Nonlinear ML models via linearization: Kernel Methods in the Proportional Regime
   - LLN-type distance-based kernel via Taylor expansion
   - CLT-type inner-product kernel via orthogonal polynomial

Two ways to linearize nonlinear models

Example (Nonlinear objects in two scaling regimes)

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector so that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries with zero mean and unit variance, and $\mathbf{y} \in \mathbb{R}^n$ be a deterministic vector of unit norm $\|\mathbf{y}\| = 1$; and consider the following two families of nonlinear objects of interest with a nonlinear function f acting on different regimes:

(i) **LLN regime**: $f(\|\mathbf{x}\|^2)$ and $f(\mathbf{x}^\mathsf{T}\mathbf{y})$; and

(ii) **CLT regime**: $f(\sqrt{n}(\|\mathbf{x}\|^2 - 1))$ and $f(\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y})$.

The two regimes follow from the two well-known convergence results:

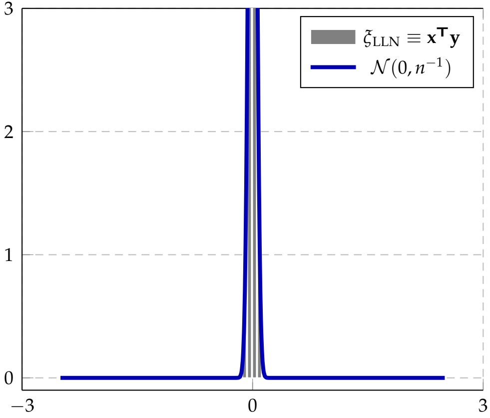
(i) **law of large numbers (LLN)**: $\|\mathbf{x}\|^2 \to \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] = 1$ and $\mathbf{x}^\mathsf{T}\mathbf{y} \to \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{y}] = 0$ almost surely as $n \to \infty$; and

(ii) **central limit theorem (CLT)**: $\sqrt{n}(\|\mathbf{x}\|^2 - 1) \to \mathcal{N}(0,2)$ and $\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \to \mathcal{N}(0,1)$ in law as $n \to \infty$.

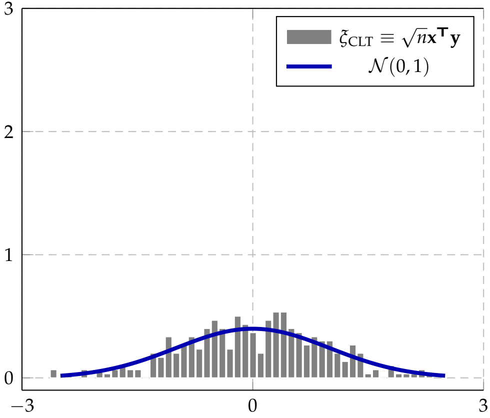
$$\|\mathbf{x}\|^2 \simeq 1 + \mathcal{N}(0,2)/\sqrt{n}, \quad \mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + \mathcal{N}(0,1)/\sqrt{n}, \tag{1}$$

for n large.

Numerical illustration



(a) LLN regime



(b) CLT regime

Figure: Illustrations of random variables in LLN (**left**) and CLT (**right**) regime, with $n = 500$.

Two different linearization techniques

$$\textbf{LLN regime } f(\|\mathbf{x}\|^2) \text{ and } f(\mathbf{x}^\mathsf{T}\mathbf{y}) \text{ versus } \textbf{CLT regime } f(\sqrt{n}(\|\mathbf{x}\|^2 - 1)) \text{ and } f(\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y})$$

two “scalings” are different:

- for objects in the LLN regime, the nonlinear function f applies on a close-to-deterministic quantity, in the sense that $\|\mathbf{x}\|^2 = 1 + O(n^{-1/2})$ and $\mathbf{x}^\mathsf{T}\mathbf{y} = 0 + O(n^{-1/2})$ with high probability for n large, due to the dominant LLN behavior; and
- for objects in the CLT regime, the nonlinear f applies on a normally distributed random variable (as a consequence of the CLT) that is not close to a deterministic quantity
- two **different** linearization approaches—via Taylor expansion and via orthogonal polynomial

Table: Comparison between two different linearization approaches.

Scaling law	LLN type	CLT type
Object of interest	$f(x)$ for (almost) deterministic $x = \tau + o(1)$	$f(x)$ for random x, e.g., $x \sim \mathcal{N}(0,1)$
Linearization technique	Taylor expansion	Orthogonal polynomial
Smoothness of f	Locally smooth f	Possibly non-smooth f

Taylor expansion

- Taylor expansion: local linearization of a smooth nonlinear function

Theorem (Taylor's theorem)

Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is at least k times continuously differentiable in a neighborhood of a given point $\tau \in \mathbb{R}$. Then, there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(\tau) + f'(x-\tau) + \frac{f''(\tau)}{2}(x-\tau)^2 + \ldots + \frac{f^{(k)}(\tau)}{k!}(x-\tau)^k + h_k(x)(x-\tau)^k, \tag{2}$$

with $\lim_{x\to\tau} h_k(x) = 0$ *so that* $h_k(x)(x-\tau)^k = o(|x-\tau|^k)$ *as* $x \to \tau$.

Working assumptions:

(i) the nonlinear function f under study should be smooth, at least in the neighborhood of the point τ of interest, so that the derivatives $f'(\tau), f''(\tau), \ldots$ make sense; and

(ii) the variable of interest x is sufficiently close to (or, concentrate around when being random) the point τ so that the higher orders terms are **neglectable**

Proposition (Taylor expansion in the LLN regime)

For random variable $x = \|\mathbf{x}\|^2$ with $\sqrt{n}\mathbf{x} \in \mathbb{R}^n$ having i.i.d. standard Gaussian entries, in the LLN regime as in Item (i) of Theorem 1, it follows from LLN and CLT that $\|\mathbf{x}\|^2 - 1 = O(n^{-1/2})$ with high probability for n large, so that one can apply Theorem 2 to write

$$f(\|\mathbf{x}\|^2) = f(1) + f'(1) \underbrace{(\|\mathbf{x}\|^2 - 1)}_{O(n^{-1/2})} + \frac{1}{2} f''(1) \underbrace{(\|\mathbf{x}\|^2 - 1)^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{3}$$

with high probability; and similarly

$$f(\mathbf{x}^\mathsf{T}\mathbf{y}) = f(0) + f'(0) \underbrace{\mathbf{x}^\mathsf{T}\mathbf{y}}_{O(n^{-1/2})} + \frac{1}{2} f''(0) \underbrace{(\mathbf{x}^\mathsf{T}\mathbf{y})^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{4}$$

again as a consequence of $\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \xrightarrow{d} \mathcal{N}(0,1)$ in distribution as $n \to \infty$, where the orders $O(n^{-\ell})$ hold with high probability for n large.

Smoothness assumption

- smoothness assumption in Taylor theorem can be relaxed
- for a non-smooth nonlinear f, can evaluate *expected* behavior $\mathbb{E}[f(x)]$ of $f(x)$, for random x
- while the function f may not be differentiable everywhere (and in particular, in the neighborhood $x = \tau$ of interest), it can still have almost everywhere **weak derivative** f' such that

$$\int f'(t)\mu(dt) = \mathbb{E}[f'(x)] < \infty, \tag{5}$$

  exists, for random variable x having law μ.
- concrete example in the case of standard Gaussian x, known in the literature as the Stein's lemma.

Lemma (Stein's lemma)

For standard Gaussian random variable $x \sim \mathcal{N}(0,1)$, we have that

$$\mathbb{E}[f'(x)] = \mathbb{E}[xf(x)], \tag{6}$$

as long as the right-hand-side term is finite.

Concentration assumption

- "closeness" or "concentration" assumption, this is a more **intrinsic limitation** of the Taylor expansion approach
- assess only the local behavior of the nonlinear function $f(x)$ around some $x = \tau$
- **otherwise**, higher-orders terms cannot be ignored (at least with high probability)
- in the CLT regime $f(\sqrt{n}(\|\mathbf{x}\|^2 - 1))$ and $f(\sqrt{n} \cdot \mathbf{x}^\mathsf{T} \mathbf{y})$, f is applied on (asymptotically) Gaussian random variables that, in particular, do not "concentrate" around any deterministic quantity

we discuss next alternative **orthogonal polynomial** approach that allows one to characterize the behavior of the nonlinear function $\mathbb{E}[f(x)]$ of random variable x that, in particular, does not strongly concentrate around a point of interest τ, as in the case of CLT regime

Motivation for orthogonal polynomial

- nonlinear function f applied on a Gaussian random variable $x \sim \mathcal{N}(0,1)$ cannot be linearized using Taylor expansion technique
- orthogonal polynomial approach can be used to "linearize" $\mathbb{E}[f(x)]$ for random and non-concentrated x, say $x \sim \mathcal{N}(0,1)$
- a functional perspective: For a random variable x of some law μ, the expectation $\mathbb{E}[f(x)]$ of the nonlinear transformation $f(x)$ for some nonlinear function f writes

$$\mathbb{E}[f(x)] = \int f(t)\mu(dt), \tag{7}$$

  for some f living in some space of functions (or, some infinite-dimensional functional space)
- Euclidean space: canonical vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ form an orthonormal basis of $\mathbb{R}^n$, so that any vector $\mathbf{x}$ living in the Euclidean space $\mathbb{R}^n$ can be decomposed as

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x}^\mathsf{T} \mathbf{e}_i)\mathbf{e}_i = \sum_{i=1}^{n} x_i \mathbf{e}_i, \tag{8}$$

  with the inner product $\mathbf{x}^\mathsf{T}\mathbf{e}_i = x_i$ the ith coordinate of $\mathbf{x}$
- a decomposition of f living in some space of functions exists: such f can be decomposed into the sum of "orthonormal" basis functions weighted by the projection of f onto these basis functions

Orthogonal polynomial

Definition (Orthogonal polynomial)

For a probability measure μ, define the inner product

$$\langle f, g \rangle \equiv \int f(x) g(x) \mu(dx) = \mathbb{E}[f(x) g(x)], \tag{9}$$

for $x \sim \mu$, we say $\{P_\ell(x), \ell \geq 0\}$ is a family of orthogonal polynomial with respect to such inner product, obtained by the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$, with $P_0(x) = 1$, P_ℓ is a polynomial function of degree ℓ and satisfies

$$\langle P_{\ell_1}, P_{\ell_2} \rangle = \mathbb{E}[P_{\ell_1}(x) P_{\ell_2}(x)] = \delta_{\ell_1 = \ell_2}. \tag{10}$$

- if the family of orthogonal polynomial $\{P_\ell(x)\}_{\ell=0}^{\infty}$ forms a orthonormal basis of $L^2(\mu)$, the set of square-integrable functions with respect to $\langle \cdot, \cdot \rangle$, any function $f \in L^2(\mu)$ can be formally expanded f

$$f(x) \sim \sum_{\ell=0}^{\infty} a_\ell P_\ell(x), \quad a_\ell = \int f(x) P_\ell(x) \mu(dx) \tag{11}$$

where "$f \sim \sum_{l=0}^{\infty} a_\ell P_\ell$" denotes that $\|f - \sum_{\ell=0}^{L} a_\ell P_\ell\|_\mu \to 0$ as $L \to \infty$ with $\|f\|_\mu^2 = \langle f, f \rangle$, or equivalently $\int \left(f(x) - \sum_{\ell=0}^{L} a_\ell P_\ell(x)\right)^2 \mu(dx) = \mathbb{E}\left[\left(f(x) - \sum_{\ell=0}^{L} a_\ell P_\ell(x)\right)^2\right] \to 0.$

Hermite polynomial

Theorem (Hermite polynomial decomposition)

For $x \in \mathbb{R}$, the ℓ^{th} order normalized Hermite polynomial, denoted $P_\ell(x)$, is given by given by

$$P_0(x) = 1, \text{ and } P_\ell(x) = \frac{(-1)^\ell}{\sqrt{\ell!}} e^{\frac{x^2}{2}} \frac{d^\ell}{dx^\ell} \left(e^{-\frac{x^2}{2}} \right), \text{ for } \ell \geq 1. \tag{12}$$

and the family of (normalized) Hermite polynomials

(i) *being orthogonal polynomials and (as the name implies) are orthonormal with respect the standard Gaussian measure, in the sense that $\int P_m(x) P_n(x) \mu(dx) = \delta_{nm}$, for $\mu(dx) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ the standard Gaussian measure*

(ii) *form an orthonormal basis of the Hilbert space (denoted $L^2(\mu)$) consist of all square-integrable functions with respect to the inner product $\langle f, g \rangle \equiv \int f(x) g(x) \mu(dx)$, and that one can formally expand any $f \in L^2(\mu)$ as*

$$f(\xi) \sim \sum_{\ell=0}^{\infty} a_{\ell,f} P_\ell(\xi), \quad a_{\ell,f} = \int f(x) P_\ell(x) \mu(dx) = \mathbb{E}[f(\xi) P_\ell(x)], \tag{13}$$

for standard Gaussian random variable $\xi \sim \mathcal{N}(0,1)$. We have

$$a_{0,f} = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[f(\xi)], \quad a_{1,f} = \mathbb{E}[\xi f(\xi)], \quad \sqrt{2} a_{2,f} = \mathbb{E}[\xi^2 f(\xi)] - a_{0,f}, \quad \nu_f = \mathbb{E}[f^2(\xi)] = \sum_{\ell=0} a_{\ell,f}^2. \tag{14}$$

Illustration of Hermite polynomial

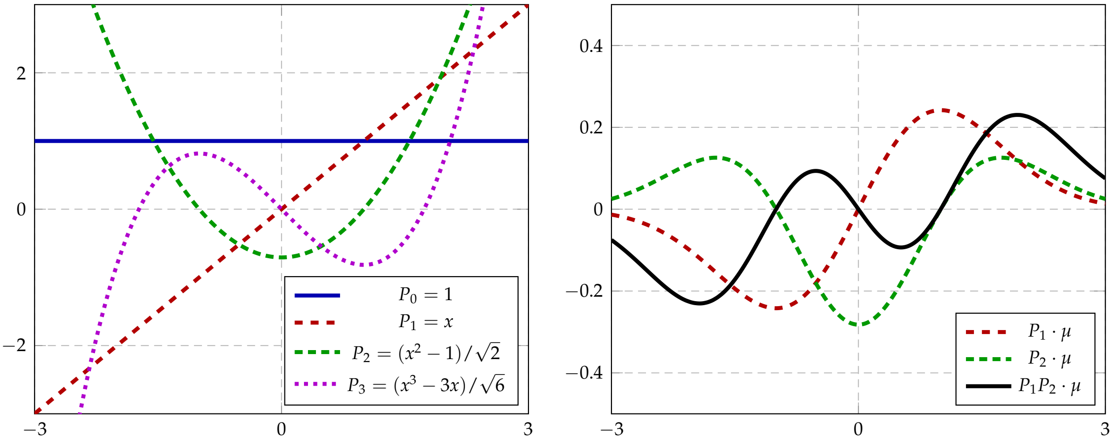


Figure: Illustration of the first four Hermite polynomials as in Theorem 5 (**left**) and of the first- and second-order Hermite polynomial (P_1 and P_2) weighted by the Gaussian mixture $\mu(dx) = \exp(-x^2/2)/\sqrt{2\pi}$ (**right**).

Different scalings, Taylor expansion versus orthogonal polynomial

For random vector $\mathbf{x} \in \mathbb{R}^n$ such that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries and deterministic $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\|_2 = 1$, $\mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0, n^{-1})$ so that

$$\xi_{\rm LLN} \equiv \mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + O(n^{-1/2}), \quad \xi_{\rm CLT} \equiv \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0,1). \tag{15}$$

We are interested in the behavior of $f(\xi_{\rm LLN})$ and $f(\xi_{\rm CLT})$:

(i) **in the LLN regime**: by Taylor expansion that any pair of smooth function f, g with $\boxed{f(0) = g(0)}$ satisfies

$$f(\xi_{\rm LLN}) = g(\xi_{\rm LLN}) + O(n^{-1/2}), \tag{16}$$

with high probability for n large, so that the two random variables $f(\xi_{\rm LLN})$ and $g(\xi_{\rm LLN})$ are close as long as the two nonlinear functions f and g coincide at 0; and

(ii) **in the CLT regime**: by Hermite polynomial decomposition that for f, g having the same zeroth-order Hermite coefficient $\boxed{a_0 = \mathbb{E}[f(\xi)] = \mathbb{E}[g(\xi)]}$ with $\xi \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[f(\xi_{\rm CLT})] = \mathbb{E}[g(\xi_{\rm CLT})]. \tag{17}$$

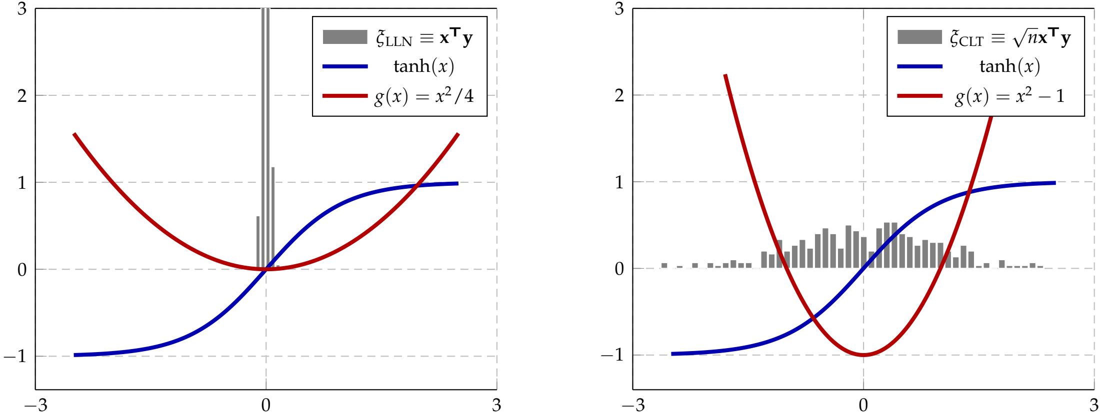
- while this is by no means surprising (by definition), orthogonal polynomials applies other nonlinear forms beyond the simple expectation $\mathbb{E}[f(\xi)]$, to nonlinear random matrix model

Example (Nonlinear behaviors of tanh in two scaling regimes)

The function $f(t) = \tanh(t)$ is "close" to *different* quadratic functions in *different* regimes of interest:

(i) **in the LLN regime**, we have $\tanh(\xi_{\rm LLN}) \simeq g(\xi_{\rm LLN})$ (so in particular $\mathbb{E}[\tanh(\xi_{\rm LLN})] \simeq \mathbb{E}[g(\xi_{\rm LLN})]$) with $g(t) = t^2/4$ as a consequence of $\tanh(x) = g(x) = 0$; and

(ii) **in the CLT regime**, we have $\mathbb{E}[\tanh(\xi_{\rm LLN})] = \mathbb{E}[g(\xi_{\rm LLN})]$ in expectation with now $g(x) = x^2 - 1$ as a consequence of the fact that their zeroth-order Hermite $a_0 = 0$.



(a) LLN regime

(b) CLT regime

Figure: Different behavior of nonlinear $f(\xi_{\mathrm{LLN}})$ and $f(\xi_{\mathrm{CLT}})$ for $f(t) = \tanh(t)$ in the LLN and CLT regime, with $n = 500$. We have in particular $\tanh(\xi_{\mathrm{LLN}}) \simeq g(\xi_{\mathrm{LLN}})$ in the LLN regime and $\mathbb{E}[\tanh(\xi_{\mathrm{CLT}})] = \mathbb{E}[g(\xi_{\mathrm{CLT}})]$ in the CLT regime with *different* g.

Take-away of this section

- two linearization techniques to linearize nonlinear objects

1. Taylor expansion: for smooth *and* concentrated objects (e.g., in the LLN regime)
2. Orthogonal polynomial approach: for non-smooth *and* non-concentrated objects (e.g., in the CLT regime)

- example: $\tanh(\xi)$ for $\xi = \xi_{\rm LLN}$ or $\xi_{\rm CLT}$ leads to **different** linearizations

Kernel matrices and their linearization

Kernel matrices: for data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, $\mathbf{K} = \{\kappa(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}$ for some $\kappa \colon \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ describe the "similarity" between data vectors.

Table: Commonly used kernels and the corresponding linearization techniques.

Family of kernel	Commonly used examples	Regime	Linearization technique
LLN-type distance-based kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = f(\\|\mathbf{x}_i - \mathbf{x}_j\\|^2/p)$	Gaussian $\exp\left(-\\|\mathbf{x}_i - \mathbf{x}_j\\|^2/(2\sigma^2 p)\right)$ Laplacian $\exp\left(-\\|\mathbf{x}_i - \mathbf{x}_j\\|/(\sigma\sqrt{p})\right)$ for some $\sigma > 0$ as well as Matérn kernel	LLN	Taylor expansion
LLN-type inner-product kernel	Polynomial $(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)^d$ for some $d \geq 1$ Sigmoid $\tanh(\beta\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ for some $\beta > 0$	LLN	Taylor expansion
CLT-type inner-product kernel	Polynomial $(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})^d$ for some $d \geq 1$ Sigmoid $\tanh(\beta\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ for some $\beta > 0$	CLT	Orthogonal polynomial

LLN-type distance-based kernel: setup

- non-trivial classification of binary GMM ($\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$ versus $\mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2)$)

$$\|\Delta\boldsymbol{\mu}\| = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = \Theta(1), \quad \|\Delta\mathbf{C}\|_2 = \|\mathbf{C}_1 - \mathbf{C}_2\|_2 = \Theta(p^{-1/2}), \tag{18}$$

- data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ extracted from a few-class (say two-class) mixture model tend to be (in the first order, and as a consequence of the LNNs) at roughly equal Euclidean distance from one another, irrespective of their corresponding class. Roughly said, in this non-trivial setting, we have

$$\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \to 0 \tag{19}$$

holds for some constant $\tau > 0$ as $n, p \to \infty$, independently of the classes, and thus of the distributions (being the same or different) of $\mathbf{x}_i$ and $\mathbf{x}_j$.

Definition (LLN-type shift-invariant kernel)

For n data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ of dimension p, we say, for *smooth* nonlinear *kernel function* $f : \mathbb{R} \to \mathbb{R}$ that

$$[\mathbf{K}]_{ij} = f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2 / p \right) \in \mathbb{R}^{n \times n}, \tag{20}$$

is a *shift-invariant* kernel matrix of the data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$. In particular, one gets the popular Gaussian kernel with $f(t) = \exp(-t/2)$.

Theorem (LLN-type shift-invariant kernel matrices via Taylor expansion, [CBG16])

Consider the non-trivial GMM classification, let $f\colon \mathbb{R} \to \mathbb{R}$ *be at least three-times differentiable in a neighborhood of* $\tau = 2 \operatorname{tr} \mathbf{C}^\circ / p = \operatorname{tr}(\mathbf{C}_1 + \mathbf{C}_2)/p$. *For a shift-invariant kernel matrix* $\mathbf{K}$, *and* $\tilde{\mathbf{K}}$ *defined below, as* $p, n \to \infty$ *with* $p/n \to c \in (0, \infty)$ *we have that* $\|\mathbf{K} - \tilde{\mathbf{K}}\|_2 = O(n^{-1/2})$. *Here,* $\tilde{\mathbf{K}}$ *is defined as*

$$\tilde{\mathbf{K}} = (f(\tau) - \tau f'(\tau)) \underbrace{\mathbf{1}_n \mathbf{1}_n^\mathsf{T}}_{\text{zeroth order}} + f'(\tau) \underbrace{\mathbf{E}}_{\text{first order}} + \frac{f''(\tau)}{2} \left(\underbrace{\boldsymbol{\psi}^2 \mathbf{1}_n^\mathsf{T} + \mathbf{1}_n (\boldsymbol{\psi}^2)^\mathsf{T} + 2 \boldsymbol{\psi} \boldsymbol{\psi}^\mathsf{T}}_{\text{second order}} \right)$$

$$+ \frac{f''(\tau)}{2} \left(\underbrace{\frac{2}{\sqrt{p}} \{(\psi_i + \psi_j)(t_a + t_b)\}_{i \neq j} + \frac{1}{p} \mathbf{J} \left(\{(t_a + t_b)^2\}_{a,b=1}^2 + 4\mathbf{T} \right) \mathbf{J}^\mathsf{T}}_{\text{second order}} \right) + (f(0) - f(\tau) + \tau f'(\tau)) \mathbf{I}_n,$$

where we denote $\mathbf{E} \in \mathbb{R}^{n \times n}$ *the (linear) Euclidean distance matrix.*

- random vector $\boldsymbol{\psi} = \{\psi_i\}_{i=1}^n \in \mathbb{R}^n$ as,

$$\psi_i \equiv \mathbf{z}_i^\mathsf{T} \mathbf{C}_a \mathbf{z}_i / p - \operatorname{tr} \mathbf{C}_a / p, \quad \text{for} \quad \mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{C}_a^{\frac{1}{2}} \mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a), \quad a \in \{1, 2\}, \tag{21}$$

- random matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, and

$$\mathbf{J} \equiv [\mathbf{j}_1, \ldots \mathbf{j}_K] \in \mathbb{R}^{n \times 2}, \tag{22}$$

- and

$$\mathbf{t} \equiv \{t_a\}_{a=1}^2 = \left\{ \frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \right\}_{a=1}^2 \in \mathbb{R}^2, \quad \mathbf{T} = \{T_{ab}\}_{a,b=1}^2 = \left\{ \frac{1}{p} \operatorname{tr} \mathbf{C}_a \mathbf{C}_b \right\}_{a,b=1}^2 \in \mathbb{R}^{2 \times 2}, \tag{23}$$

with $\mathbf{j}_a \in \mathbb{R}^n$ the canonical vector of class $\mathcal{C}_a$, that is, $[\mathbf{j}_a]_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$; and $\mathbf{t}, \mathbf{T}$ functions of the data covariances $\mathbf{C}_1, \mathbf{C}_2$.

Proof

- expansion of "normalized" Euclidean distance:

$$\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2=\underbrace{\frac{2}{p}\operatorname{tr}\mathbf{C}^\circ}_{\equiv\tau=O(1)}+\underbrace{\psi_i+\psi_j+\frac{1}{p}\operatorname{tr}(\mathbf{C}_a^\circ+\mathbf{C}_b^\circ)-\frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a^{\frac{1}{2}}\mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j}_{O(p^{-1/2})}$$
$$+\underbrace{\frac{1}{p}\|\boldsymbol{\mu}_a-\boldsymbol{\mu}_b\|^2+\frac{2}{p}(\boldsymbol{\mu}_a-\boldsymbol{\mu}_b)^\mathsf{T}(\mathbf{C}_a^{\frac{1}{2}}\mathbf{z}_i-\mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j)}_{O(p^{-1})}, \tag{24}$$

with $\mathbf{C}^\circ\equiv\frac{1}{2}(\mathbf{C}_1+\mathbf{C}_2)$ the centered covariance and $\mathbf{C}_a^\circ\equiv\mathbf{C}_a-\mathbf{C}^\circ$ so that $\|\mathbf{C}_a^\circ\|_2=\frac{1}{2}\|\Delta\mathbf{C}\|_2=O(p^{-1/2})$, as well as $\psi_i\equiv\mathbf{z}_i^\mathsf{T}\mathbf{C}_a\mathbf{z}_i/p-\operatorname{tr}\mathbf{C}_a/p=O(p^{-1/2})$.

- Taylor-expanding $[\mathbf{K}]_{ij}=f(\|\mathbf{x}_i-\mathbf{x}_j\|^2/p)$ around $f(\tau)$ that

$$[\mathbf{K}]_{ij}=f\left(\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2\right)$$
$$=\underbrace{f(\tau)}_{\equiv K_0=O(1)}+\underbrace{f'(\tau)\left(\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right)}_{\equiv K_1=O(p^{-1/2})}+\underbrace{\frac{1}{2}f''(\tau)\left(\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right)^2}_{\equiv K_2=O(p^{-1})}+\underbrace{O(p^{-3/2})}_{\equiv K_3}, \tag{25}$$

Proof

- by $\|\mathbf{A}\|_2 \leq n\|\mathbf{A}\|_\infty$ for matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, we know that the higher-order terms $O(p^{-3/2})$, when put in matrix form, are of spectral norm order $O(n^{-1/2})$ and thus vanish asymptotically as $n, p \to \infty$.

(i) the leading order term is $K_0 = f(\tau) = O(1)$ and, as in the case of Euclidean distance matrix in the linear case, does not depend on the data $\mathbf{x}_i, \mathbf{x}_j$ (or their classes); and

(ii) the second-order term K_1 is proportional to $f'(\tau)$, of order $O(p^{-1/2})$, is the same as in the linear Euclidean distance matrix $\mathbf{E}$ with $f(t) = t$; and

(iii) the third-order term K_2 is proportional to $f''(\tau)$, of order $O(p^{-1})$, contains quadratic function of $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ and therefore crucially differs from the linear $f(t) = t$ scenario.

- the (i, j) entry of the nonlinear kernel matrix $\mathbf{K}$ takes a similar form as the linear Euclidean distance matrix $\mathbf{E}$ (with $f(t) = t$), but with a few additional and nonlinear terms collected in K_2 that are proportional to $f''(\tau)$.

Additional nonlinear terms: only the terms of order $O(n^{-1/2})$ in $\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau$ will remain after taking the square, that is

$$\left(\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau\right)^2 = \left(\psi_i + \psi_j + \frac{1}{p}\operatorname{tr}(\mathbf{C}_a^\circ + \mathbf{C}_b^\circ) - \frac{2}{p}\mathbf{z}_i^\mathsf{T}\mathbf{C}_a^{\frac{1}{2}}\mathbf{C}_b^{\frac{1}{2}}\mathbf{z}_j\right)^2 + O(n^{-3/2}) \tag{26}$$

Proof

This, in matrix form (with $i \neq j$ for the moment),

$$\left\{ \left(\frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right)^2 \right\}_{i \neq j} = \left\{ \left(\psi_i + \psi_j + \frac{1}{p} \operatorname{tr}(\mathbf{C}_a^\circ + \mathbf{C}_b^\circ) \right)^2 + 4 \left(\frac{1}{p} \mathbf{z}_i^\mathsf{T} \mathbf{C}_a^{\frac{1}{2}} \mathbf{C}_b^{\frac{1}{2}} \mathbf{z}_j \right)^2 \right\}_{i \neq j}$$
$$- \left\{ \frac{4}{p} \mathbf{z}_i^\mathsf{T} \mathbf{C}_a^{\frac{1}{2}} \mathbf{C}_b^{\frac{1}{2}} \mathbf{z}_j \left(\psi_i + \psi_j + \frac{1}{p} \operatorname{tr}(\mathbf{C}_a^\circ + \mathbf{C}_b^\circ) \right) \right\}_{i \neq j} + O_{\|\cdot\|}(n^{-1/2})$$
$$= \boldsymbol{\psi}^2 \mathbf{1}_n^\mathsf{T} + \mathbf{1}_n (\boldsymbol{\psi}^2)^\mathsf{T} + 2 \boldsymbol{\psi} \boldsymbol{\psi}^\mathsf{T} + \frac{2}{\sqrt{p}} \{ (\psi_i + \psi_j)(t_a + t_b) \}_{i \neq j}$$
$$+ \frac{1}{p} \mathbf{J} \left(\{ (t_a + t_b)^2 \}_{a,b=1}^2 + 4\mathbf{T} \right) \mathbf{J}^\mathsf{T} + O_{\|\cdot\|}(n^{-1/2}), \tag{27}$$

where we denote $\boldsymbol{\psi}^2 \equiv \{\psi_i^2\}_{i=1}^n \in \mathbb{R}^n$, $O_{\|\cdot\|}(n^{-1/2})$ for matrices of spectral norm ($\|\cdot\|$) order $O(n^{-1/2})$, and

$$\left\{ \left(\frac{1}{p} \mathbf{z}_i^\mathsf{T} \mathbf{C}_a^{\frac{1}{2}} \mathbf{C}_b^{\frac{1}{2}} \mathbf{z}_j \right)^2 \right\}_{i \neq j} = \left\{ \mathbb{E} \left(\frac{1}{p} \mathbf{z}_i^\mathsf{T} \mathbf{C}_a^{\frac{1}{2}} \mathbf{C}_b^{\frac{1}{2}} \mathbf{z}_j \right)^2 \right\}_{i \neq j} + O_{\|\cdot\|}(n^{-1/2})$$
$$= \left\{ \frac{1}{p^2} \operatorname{tr} \mathbf{C}_a \mathbf{C}_b \right\}_{i \neq j} + O_{\|\cdot\|}(n^{-1/2}) \equiv \frac{1}{p} \mathbf{J} \mathbf{T} \mathbf{J}^\mathsf{T} + O_{\|\cdot\|}(n^{-1/2}), \tag{28}$$

Discussions

- "linearizes" the nonlinear kernel matrix $\mathbf{K}$ for smooth kernel function f, and see both linear terms $\mathbf{E}$ (K_0 and K_1) and higher-order nonlinear terms K_2 in the linearization $\tilde{\mathbf{K}}$

(i) it follows from the derivations in Equation (27) and Equation (28) that the higher-order nonlinear terms in $\tilde{\mathbf{K}}$ are approximately (in a spectral norm sense) of low rank, for n, p large; and

(ii) as a consequence, the eigenspectrum of $\tilde{\mathbf{K}}$ (and thus of $\mathbf{K}$ by Theorem 8) is like that of the Euclidean distance matrix $\mathbf{E}$, scaled by $f'(\tau)$, and with a few additional spiked eigenvalues due to the higher-order nonlinear terms in K_2.

Theorem (Limiting spectrum of shift-invariant kernel matrices)

Under the same assumptions and notations of Theorem 8, we have, for $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$, $f'(\tau) \neq 0$, and as $p, n \to \infty$ with $p/n \to c \in (0, \infty)$, that the empirical spectral measure of the shift-invariant kernel matrix $\mathbf{K}$ converges weakly and almost surely to the rescaled and shifted Marčenko–Pastur law $-2f'(\tau)\mu_{\mathrm{MP},c^{-1}} + \kappa$, $\kappa = f(0) - f(\tau) + \tau f'(\tau)$, which is the law of $-2f'(\tau)x + \kappa$ for x following a Marčenko–Pastur distribution with parameter c^{-1}, i.e., $x \sim \mu_{\mathrm{MP},c^{-1}}$.

Numerical results

- $f_1(t) = \exp(-t/2)$, that corresponds to the Gaussian kernel matrix
- versus $f_2(t) = at^2 + bt + c$, that corresponds to the polynomial kernel matrix, where the parameters a, b, and c are chosen such that

$$a = \frac{1}{8}\exp(-\tau/2), \quad b = -\frac{1}{2}\exp(-\tau/2) - \frac{\tau}{4}\exp(-\tau/2), \quad c = \exp(-\tau/2) - a\tau^2 - b\tau. \tag{29}$$

- the two functions share the same values of $f(\tau), f'(\tau), f''(\tau)$, i.e., they have the same local behavior per Taylor expansion

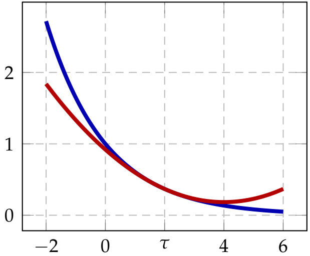
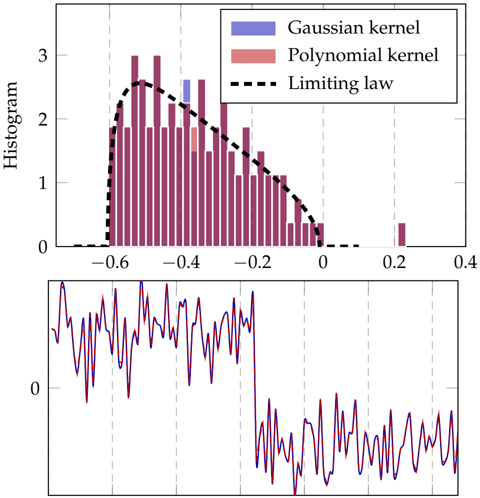
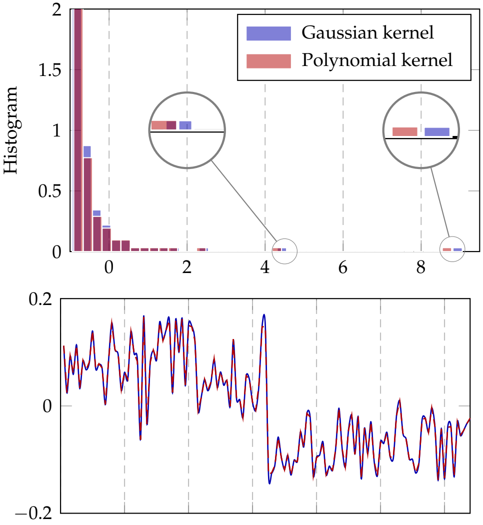


Figure: Different kernel function $f_1(t) = \exp(-t/2)$ versus polynomial $f_2(t)$ given in Equation (29), with similar local behavior around $\tau = 2$.



(a) Gaussian mixture data



(b) MNIST data (number 0 versus 1)

CLT-type inner-product kernel matrix: setup

Definition (CLT-type inner-product kernel)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n data vectors of dimension p, and let $f : \mathbb{R} \to \mathbb{R}$ be a possibly *non-smooth* nonlinear *kernel function* (that is square integrable to standard Gaussian measure). Then, we say that

$$[\mathbf{K}]_{ij} = \begin{cases} f(\mathbf{x}_i^\top \mathbf{x}_j / \sqrt{p}) / \sqrt{p} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \tag{30}$$

is a CLT-type inner-product kernel matrix for i.i.d. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. In this case, we denote, as in Equation Equation (14), the Hermite coefficients of f as

$$a_0 = \mathbb{E}[f(\xi)], \quad a_1 = \mathbb{E}[\xi f(\xi)], \quad \nu = \mathbb{E}[f^2(\xi)], \tag{31}$$

for $\xi \sim \mathcal{N}(0, 1)$. Without loss of generality, we assume the nonlinear kernel function f is "centered" with respect to standard Gaussian measure with $a_0 = 0$ (which can be achieved by studying $\tilde{f}(x) = f(x) - \mathbb{E}[f(\xi)]$).

Theorem (Limiting spectrum of CLT-type inner-product kernel matrices, [CS13; DV13])

Let $p, n \to \infty$ with $p/n \to c \in (0, \infty)$ and assume $f : \mathbb{R} \to \mathbb{R}$ is square-integrable with respect to standard Gaussian measure with $a_0 = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)}[f(\xi)] = 0$. Then, the empirical spectral measure of the inner-product kernel matrix $\mathbf{K}$ defined in Theorem 10 converges weakly and almost surely to a probability measure μ defined by its Stieltjes transform $m(z)$, as the unique solution to

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z), \tag{32}$$

for a_1, ν the Hermite coefficients of f defined in Equation (31).

Theorem (A matrix version of asymptotic equivalent linear model)

Under the same settings above, when the limiting spectral measure is considered, the inner-product random kernel matrix $\mathbf{K}$ admits the following asymptotic equivalent linear model,

$$\mathbf{K} \equiv f(\mathbf{X}^\mathsf{T}\mathbf{X}/\sqrt{p})/\sqrt{p} - \operatorname{diag}(\cdot) \leftrightarrow \tilde{\mathbf{K}}_f = a_1 \mathbf{X}^\mathsf{T}\mathbf{X}/p + \sqrt{\nu - a_1^2} \cdot \mathbf{Z}/\sqrt{p} - \operatorname{diag}(\cdot), \tag{33}$$

where we use $\mathbf{A} - \operatorname{diag}(\mathbf{A})$ to get a matrix with zeros on its diagonal, and with its non-diagonal entries same as $\mathbf{A}$.

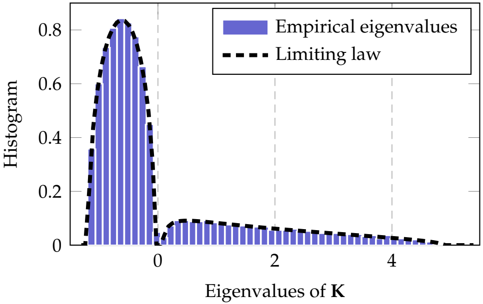
Remark

- As a consequence of the form of $m(z)$, the limiting spectral measure μ of $\mathbf{K}$ is the free additive convolution (denoted as '$\boxplus$', see [VDN92; Bia98] for an introduction) between the Marčenko–Pastur law (denoted $\mu_{\mathrm{MP},c}$ of shape parameter $c = \lim p/n$) and the so-called Wigner semicircle law (denoted μ_{SC}) as

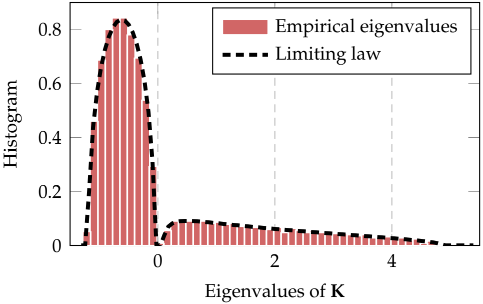
$$\mu = a_1(\mu_{\mathrm{MP},c^{-1}} - 1) \boxplus \sqrt{(\nu - a_1^2)c^{-1}}\mu_{\mathrm{SC}}, \tag{34}$$

  where $a_1(\mu_{\mathrm{MP},c^{-1}} - 1)$ is the law of $a_1(x-1)$ for $x \sim \mu_{\mathrm{MP},c^{-1}}$ and $\sqrt{(\nu - a_1^2)c^{-1}}\mu_{\mathrm{SC}}$ the law of $\sqrt{(\nu - a_1^2)c^{-1}} \cdot x$ for $x \sim \mu_{\mathrm{SC}}$.
- intuitively, the Marčenko–Pastur law characterizes the linear part ($a_1 x$) of the nonlinear kernel function $f(x)$, while the higher-order "purely" nonlinear part $f(x) - a_1 x$ contributes to the semicircle law.
- these two contributions are asymptotically "independent" so that the resulting limiting spectrum is the free additive convolution of each component.

Numerical results



(a) $f_1(t) = \tanh(t)$



(b) quadratic $f_2(t) = 0.1171(t^2 - 1) + 0.6057t$

Figure: Eigenvalues of inner-product kernel matrices $\mathbf{K}$ defined in Equation (30) for different nonlinear kernel functions f_1 and f_2, versus the limiting law given in Theorem 11, for $p = 512$, $n = 2\,048$, $f_1(t) = \tanh(t)$ versus quadratic $f_2(t)$ that share the same parameters of a_1 and ν.

Numerical results

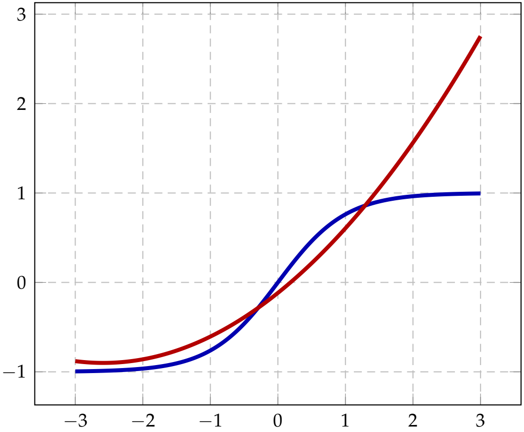


Figure: Different kernel function $f_1(t) = \tanh(t)$ versus polynomial $f_2(t) = 0.1171(t^2 - 1) + 0.6057t$ that lead to asymptotically similar kernel eigenspectral behavior. In particular, this figure is to be compared with Figure 4, where we observe a (Taylor-expansion) concentration point in the latter. Here, the two nonlinear functions f_1 and f_2 are *not* locally close (e.g., in the sense of Taylor expansion), but only share the same Hermite coefficients a_1 and ν.

Take-away messages of this section

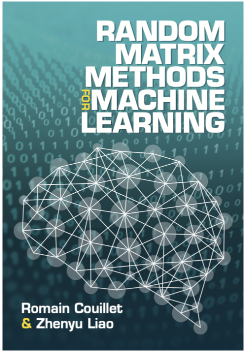
- linearization of nonlinear kernel matrices $\mathbf{K}$

1. **LLN-type** nonlinear kernel matrices: Taylor expansion
2. **CLT-type** nonlinear kernel matrices: Orthogonal polynomial

- **local** versus **global** perspective of the non-linearity

Random matrix theory (RMT) for machine learning:

- **change of intuition** from small to large dimensional learning paradigm!
- **better understanding** of existing methods: why they work if they do, and what the issue is if they do not
- **improved novel methods** with performance guarantee!



- book *"Random Matrix Methods for Machine Learning"*
- by Romain Couillet and **Zhenyu Liao**
- Cambridge University Press, 2022
- a pre-production version of the book and exercise solutions at `https://zhenyu-liao.github.io/book/`
- MATLAB and Python codes to reproduce all figures at `https://github.com/Zhenyu-LIAO/RMT4ML`

Thank you! Q & A?