Kernelized Cumulants: Beyond Mean Embeddings

Zoltán Szabó

Joint work with:

- Patric Bonnier, Harald Oberhauser
- @ Mathematical Institute, University of Oxford.


Statistics Research Showcase, LSE
June 5, 2023

Moments and cumulants on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

Moments and cumulants on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

- Cumulants $\kappa(\gamma) = \left(\kappa^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$: from the moment-generating function

$$\sum_{i\in\mathbb{N}} \kappa^{(i)}(\gamma)\frac{\theta^i}{i!} = \log\left(\sum_{i\in\mathbb{N}} \mu^{(i)}(\gamma)\frac{\theta^i}{i!}\right).$$

Moments and cumulants on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

- Cumulants $\kappa(\gamma) = \left(\kappa^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$: from the moment-generating function

$$\sum_{i\in\mathbb{N}} \kappa^{(i)}(\gamma)\frac{\theta^i}{i!} = \log\left(\sum_{i\in\mathbb{N}} \mu^{(i)}(\gamma)\frac{\theta^i}{i!}\right).$$

$\kappa^{(1)}(\gamma) = \mathbb{E}(X)$	mean
$\kappa^{(2)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^2$	variance
$\kappa^{(3)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^3$	3rd central moment
$\kappa^{(4)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^4 - 3\left[\mathbb{E}(X - \mathbb{E}X)^2\right]^2$	
$\kappa^{(5)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^5 - 10\mathbb{E}(X - \mathbb{E}X)^3\mathbb{E}(X - \mathbb{E}X)^2$	

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description

$\kappa^{(1)}(\gamma) = \mathbb{E}(X)$, $\{\{1\}\}$

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X)$$

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

where $X, X' \sim \gamma$, independent.

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

$$\kappa^{(3)}(\gamma) = \mathbb{E}\left(X^3\right) - 3\mathbb{E}\left(X^2\right)\mathbb{E}(X) + 2\mathbb{E}^3(X)$$

where $X, X' \sim \gamma$, independent.

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

$$\kappa^{(3)}(\gamma) = \mathbb{E}\left(X^3\right) - \overbrace{3\mathbb{E}\left(X^2\right)\mathbb{E}(X)}^{\mathbb{E}(XXX')+\mathbb{E}(XX'X)+\mathbb{E}(X'XX)} + 2\mathbb{E}^3(X), \qquad \{\{1,2,3\}\}, \{\{1,2\},\{3\}\}, \{\{1,3\},\{2\}\}, \{\{2,3\},\{1\}\}, \{\{1\},\{2\},\{3\}\},$$

$\ldots$

where $X, X' \sim \gamma$, independent.

Question

What are the weights in front of the moments?

Unzipping cumulants on $\mathbb{R}$: the weights

m	elements of $\pi \in P(m)$	$\lvert\pi\rvert$	c_π
1	$\{1\}$	1	1
2	$\{1,2\}$	1	1
	$\{1\},\{2\}$	2	-1
3	$\{1,2,3\}$	1	1
	$\{1,2\}, \{3\}$	2	-1
	$\{1,3\}, \{2\}$	2	-1
	$\{2,3\}, \{1\}$	2	-1
	$\{1\}, \{2\}, \{3\}$	3	2

with $P(m) :=$ all partitions of $[m]$, $c_\pi = (-1)^{|\pi|-1}(|\pi|-1)!$

Motivation, i.e. one reason why one likes cumulants

Moment and cumulants on $\mathbb{R}^d$

Change $\mathbb{E}\left(X^i\right) \in \mathbb{R}$ to $\mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$ ($\mathbf{i} \in \mathbb{N}^d$). log, $P(m)$: ✓

Known theorem [Billingsley, 2012]

Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1, \ldots, X_d) \sim \gamma$. Then

1. $\gamma \mapsto \kappa(\gamma)$ is injective.
2. $X_1, \ldots, X_d$ are independent $\Leftrightarrow \kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.

Motivation, i.e. one reason why one likes cumulants

Moment and cumulants on $\mathbb{R}^d$

Change $\mathbb{E}\left(X^i\right) \in \mathbb{R}$ to $\mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$ ($\mathbf{i} \in \mathbb{N}^d$). log, $P(m)$: ✓

Known theorem [Billingsley, 2012]

Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1, \ldots, X_d) \sim \gamma$. Then

1. $\gamma \mapsto \kappa(\gamma)$ is injective.
2. $X_1, \ldots, X_d$ are independent $\Leftrightarrow \kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.

Motivation

1. Various data types, nonlinear features: kernels.
2. Linear: not even characteristic (see MMD and HSIC).
3. Computable estimators.

Idea

Lifting

$(X_1, \ldots, X_d) \in \times_{j=1}^{d} \mathcal{X}_j \to (\Phi_1(X_1), \ldots, \Phi_d(X_d)) \in \times_{j=1}^{d} \mathcal{H}_{k_j}.$

Idea

Lifting

$$(X_1, \dots, X_d) \in \times_{j=1}^d \mathcal{X}_j \to (\Phi_1(X_1), \dots, \Phi_d(X_d)) \in \times_{j=1}^d \mathcal{H}_{k_j}.$$

Ingredients:

1. Moments: swap out $\mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$ to

$$\mathbb{E}\left[[\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}\right] \in \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}.$$

Idea

Lifting

$$(X_1, \ldots, X_d) \in \times_{j=1}^d \mathcal{X}_j \to (\Phi_1(X_1), \ldots, \Phi_d(X_d)) \in \times_{j=1}^d \mathcal{H}_{k_j}.$$

Ingredients:

1. Moments: swap out $\mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$ to

$$\mathbb{E}\left[[\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}\right] \in \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}.$$

2. From moments to cumulants:
   - log on tensor algebras, or
   - combinatorial description of cumulants ($\leftarrow$ a bit simpler, but $\Leftrightarrow$).

Idea

Lifting

$(X_1, \dots, X_d) \in \times_{j=1}^d \mathcal{X}_j \to (\Phi_1(X_1), \dots, \Phi_d(X_d)) \in \times_{j=1}^d \mathcal{H}_{k_j}.$

Ingredients:

1. Moments: swap out $\mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$ to

$$\mathbb{E}\left[[\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}\right] \in \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}.$$

2. From moments to cumulants:
   - log on tensor algebras, or
   - combinatorial description of cumulants ($\leftarrow$ a bit simpler, but $\Leftrightarrow$).
3. Computation: by the 'expected kernel trick' (V-statistics).

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

- Def-1 (feature space):

$$k(a,b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

- Def-1 (feature space):

$$k(a,b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n} \succeq 0$.

Notes

- $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k = \overline{\operatorname{Span}}(k(\cdot, x) : x \in \mathcal{X})$: Fourier analysis, polynomials, splines, ...
- Examples: $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$.
- Kernels exist on various domains!

Mean embedding

- Mean embedding (Bochner integral):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

Mean embedding, MMD

- Mean embedding (Bochner integral):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

Mean embedding, MMD, HSIC

- Mean embedding (Bochner integral):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion, $k := \otimes_{j=1}^d k_j$:

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{j=1}^d \mathbb{P}_j\right)$$

Mean embedding, MMD, HSIC

- Mean embedding (Bochner integral):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion, $k := \otimes_{j=1}^{d} k_j$:

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{j=1}^{d}\mathbb{P}_j\right),$$
$$= \Big\| \underbrace{\mu_{\otimes_{j=1}^{d} k_j}(\mathbb{P}) - \otimes_{j=1}^{d}\mu_{k_j}(\mathbb{P}_j)}_{\text{cross-covariance operator}} \Big\|_{\mathcal{H}_k}.$$

Mean embedding, MMD, HSIC

- Mean embedding (Bochner integral):

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion, $k := \otimes_{j=1}^d k_j$:

$$\begin{aligned}\mathrm{HSIC}_k(\mathbb{P}) &:= \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{j=1}^d \mathbb{P}_j\right)\\ &= \Bigg\| \underbrace{\mu_{\otimes_{j=1}^d k_j}(\mathbb{P}) - \otimes_{j=1}^d \mu_{k_j}(\mathbb{P}_j)}_{\text{cross-covariance operator}} \Bigg\|_{\mathcal{H}_k}.\end{aligned}$$

Clarification of what $\otimes_{j=1}^d k_j$ and $\otimes_{j=1}^d \mu_{k_j}(\mathbb{P}_j)$ are follows.

Tensor product: $\bigotimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.

Tensor product: $\otimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $d = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

Tensor product: $\otimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{ab}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{ab}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $d = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

- For $d \geq 2$ and $a_j \in \mathcal{H}_j$,

$$\left(\otimes_{j=1}^{d} a_j\right) (b_1, \ldots, b_d) := \prod_{j=1}^{d} \langle a_j, b_j \rangle_{\mathcal{H}_j}.$$

Tensor product: $\otimes_{j=1}^{d}\mathcal{H}_j$

$$\otimes_{j=1}^{d}\mathcal{H}_j := \overline{\mathrm{Span}}(\otimes_{j=1}^{d} a_j \,:\, a_j \in \mathcal{H}_j).$$

Tensor product: $\otimes_{j=1}^d \mathcal{H}_j$

$$\otimes_{j=1}^d \mathcal{H}_j := \overline{\mathrm{Span}}(\otimes_{j=1}^d a_j \,:\, a_j \in \mathcal{H}_j).$$

$\xrightarrow{\text{spec.}}$ The tensor product of RKHSs is an RKHS

$$\mathcal{H}_k = \otimes_{j=1}^d \mathcal{H}_{k_j},$$

$$k(x, x') := (\otimes_{j=1}^d k_j)(x, x') := \prod_{j=1}^d \underbrace{k_j(x_j, x_j')}_{\text{coordinate-wise similarity}}.$$

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$: k is characteristic.

Validness of MMD & HSIC, their estimation

Validness:

- $\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$: k is characteristic.
- $\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{j=1}^{d} \mathbb{P}_j \Leftarrow k_j$-s are universal.

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$: k is characteristic.
- $\mathrm{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{j=1}^{d}\mathbb{P}_j \Leftarrow k_j$-s are universal.

Properties:

1. Injectivity of μ_k on probability / finite signed measures, so

   universal $\Rightarrow$ characteristic.

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$: k is characteristic.
- $\mathrm{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{j=1}^{d}\mathbb{P}_j \Leftarrow k_j$-s are universal.

Properties:

1. Injectivity of μ_k on probability / finite signed measures, so

   universal $\Rightarrow$ characteristic.

2. Easy-to-estimate: expected kernel trick

$$\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y).$$

Kernelized moments – towards kernelized cumulants

- From now:
  - $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$, $X \sim \gamma$,
  - kernels $k_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$, $j \in [d]$,
  - lifting $\Phi(X) = (\Phi_j(X_j))_{j=1}^d$ with $\Phi_j(x_j) := k_j(\cdot, x_j)$,
  - RKHS $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}$ with kernel $k^{\otimes \mathbf{i}} := k_1^{\otimes i_1} \otimes \cdots \otimes k_d^{\otimes i_d}$, and feature

$$\Phi^{\otimes \mathbf{i}}(X) := [\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}.$$

Kernelized moments – towards kernelized cumulants

- From now:
  - $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$, $X \sim \gamma$,
  - kernels $k_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$, $j \in [d]$,
  - lifting $\Phi(X) = (\Phi_j(X_j))_{j=1}^d$ with $\Phi_j(x_j) := k_j(\cdot, x_j)$,
  - RKHS $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}$ with kernel $k^{\otimes \mathbf{i}} := k_1^{\otimes i_1} \otimes \cdots \otimes k_d^{\otimes i_d}$, and feature

$$\Phi^{\otimes \mathbf{i}}(X) := [\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}.$$

- Moment sequence:

$$\mu(\gamma) = \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}, \qquad \mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[\Phi^{\otimes \mathbf{i}}(X)\right] \in \mathcal{H}^{\otimes \mathbf{i}}.$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X \sim \gamma$,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\left[\Phi(X)\right]$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X \sim \gamma$,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\left[\Phi(X)\right],$$
$$\kappa_k^{(2)}(\gamma) = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right] - \mathbb{E}\left[\Phi(X)\right] \otimes \mathbb{E}\left[\Phi(X)\right]$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\big[\Phi(X)\big],$$

$$\kappa_k^{(2)}(\gamma) = \mathbb{E}\big[\Phi(X) \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X)\big] \otimes \mathbb{E}\big[\Phi(X)\big],$$

$$\begin{aligned}\kappa_k^{(3)}(\gamma) = {} & \mathbb{E}\Big[\Phi^{\otimes 3}(X)\ \Big] - \mathbb{E}\big[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\big] \\ & - \mathbb{E}\big[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\big] \\ & + 2\mathbb{E}^{\otimes 3}\big[\Phi(X)\big].\end{aligned}$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\big[\Phi(X)\big],$$
$$\kappa_k^{(2)}(\gamma) = \mathbb{E}\big[\Phi(X) \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X)\big] \otimes \mathbb{E}\big[\Phi(X)\big],$$
$$\begin{aligned}\kappa_k^{(3)}(\gamma) = \mathbb{E}\Big[\Phi^{\otimes 3}(X)\ \Big] &- \mathbb{E}\big[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\big] \\ &- \mathbb{E}\big[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\big] \\ &+ 2\mathbb{E}^{\otimes 3}\big[\Phi(X)\big].\end{aligned}$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$\kappa_{k_1,k_2}^{(2,0)}(\gamma) = \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$
\begin{aligned}
\kappa_k^{(1)}(\gamma) &= \mathbb{E}\big[\Phi(X)\big],\\
\kappa_k^{(2)}(\gamma) &= \mathbb{E}\big[\Phi(X) \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X)\big] \otimes \mathbb{E}\big[\Phi(X)\big],\\
\kappa_k^{(3)}(\gamma) &= \mathbb{E}\Big[\Phi^{\otimes 3}(X)\ \Big] - \mathbb{E}\big[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\big]\\
&\quad - \mathbb{E}\big[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\big]\\
&\quad + 2\mathbb{E}^{\otimes 3}\big[\Phi(X)\big].
\end{aligned}
$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$
\begin{aligned}
\kappa_{k_1,k_2}^{(2,0)}(\gamma) &= \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],\\
\kappa_{k_1,k_2}^{(1,1)}(\gamma) &= \mathbb{E}\left[\Phi_1(X_1) \otimes \Phi_2(X_2)\right] - \mathbb{E}\left[\Phi_1(X_1)\right] \otimes \mathbb{E}\left[\Phi_2(X_2)\right]
\end{aligned}
$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}[\Phi(X)],$$
$$\kappa_k^{(2)}(\gamma) = \mathbb{E}[\Phi(X) \otimes \Phi(X)] - \mathbb{E}[\Phi(X)] \otimes \mathbb{E}[\Phi(X)],$$
$$\begin{aligned}\kappa_k^{(3)}(\gamma) = {} & \mathbb{E}\left[\Phi^{\otimes 3}(X)\right] - \mathbb{E}[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')] \\ & - \mathbb{E}[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)] - \mathbb{E}[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)] \\ & + 2\mathbb{E}^{\otimes 3}[\Phi(X)].\end{aligned}$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$\kappa_{k_1,k_2}^{(2,0)}(\gamma) = \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],$$
$$\kappa_{k_1,k_2}^{(1,1)}(\gamma) = \mathbb{E}\left[\Phi_1(X_1) \otimes \Phi_2(X_2)\right] - \mathbb{E}\left[\Phi_1(X_1)\right] \otimes \mathbb{E}\left[\Phi_2(X_2)\right],$$
$$\kappa_{k_1,k_2}^{(0,2)}(\gamma) = \mathbb{E}\left[\Phi_2^{\otimes 2}(X_2)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_2(X_2)\right].$$

Wanted: repetition and partitioning. Weights: as before (c_π).

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1,\ldots,X_1}_{i_1 \text{ times}},\underbrace{X_2,\ldots,X_2}_{i_2 \text{ times}},\ldots,\underbrace{X_d,\ldots,X_d}_{i_d \text{ times}}).$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^{d} \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^d \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Kernelized cumulants: $m = \deg(\mathbf{i}) := \sum_{j=1}^d i_j \xrightarrow{\text{OK}} \gamma_\pi^{\mathbf{i}} = (\gamma^{\mathbf{i}})_\pi$,

$$\kappa_{k_1,\ldots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d},$$
$$\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) := \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes \mathbf{i}}(\cdot, (X_1, \ldots, X_m)).$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^{d} \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Kernelized cumulants: $m = \deg(\mathbf{i}) := \sum_{j=1}^{d} i_j \overset{\text{OK}}{\Longrightarrow} \gamma_\pi^{\mathbf{i}} = \left(\gamma^{\mathbf{i}}\right)_\pi$,

$$\kappa_{k_1,\ldots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d},$$
$$\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) := \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes \mathbf{i}}(\cdot, (X_1, \ldots, X_m)).$$

$\Rightarrow$ expected kernel trick is applicable

Cumulants characterize distributions

Point-separating k := injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Cumulants characterize distributions

Point-separating k := injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).

Cumulants characterize distributions

Point-separating k := injectivity of Φ $\Leftarrow$ characteristic k $\Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \eta \Leftrightarrow \kappa_{k_1,\ldots,k_d}(\gamma) = \kappa_{k_1,\ldots,k_d}(\eta)$

Cumulants characterize distributions

Point-separating k $:=$ injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \eta \Leftrightarrow \kappa_{k_1,\dots,k_d}(\gamma) = \kappa_{k_1,\dots,k_d}(\eta)$, and

$$
\begin{aligned}
d^{\mathbf{i}}(\gamma,\eta) &:= \|\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma) - \kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\eta)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \Big[\mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\
&\qquad + \mathbb{E}_{\eta^{\mathbf{i}}_\pi \otimes \eta^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\
&\qquad - 2\mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \eta^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m))\Big].
\end{aligned}
$$

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$, and

$$\|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_j)_{j=1}^m, (Y_j)_{j=1}^m),$$

where $m = \deg(\mathbf{i})$.

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$, and

$$\|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}\left((X_j)_{j=1}^m, (Y_j)_{j=1}^m\right),$$

  where $m = \deg(\mathbf{i})$.

Estimation in both cases

$\mathbb{E}k^{\otimes \mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m)) \Rightarrow$ V-statistics ✓

Distance between kernel variance embeddings

- By our theorem: if $\gamma = \eta$, then $d^{(2)}(\gamma, \eta) = 0$.
- V-statistic estimator of $d^{(2)}(\gamma, \eta)$:

$$\frac{1}{N^2}\mathrm{Tr}\left[(\mathbf{K}_x\mathbf{J}_N)^2\right] + \frac{1}{M^2}\mathrm{Tr}\left[(\mathbf{K}_y\mathbf{J}_M)^2\right] - \frac{2}{NM}\mathrm{Tr}\left[\mathbf{K}_{xy}\mathbf{J}_M\mathbf{K}_{xy}^\top\mathbf{J}_N\right],$$

with $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \gamma$, $(y_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \eta$, $\mathbf{K}_x = [k(x_i, x_j)]_{i,j=1}^N$, $\mathbf{K}_y = [k(y_i, y_j)]_{i,j=1}^M$, $\mathbf{K}_{x,y} = [k(x_i, y_j)]_{i,j=1}^{N,M}$, $\mathbf{J}_n = \mathbf{I}_n - \mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

Distance between kernel variance/skewness embeddings

- By our theorem: if $\gamma = \eta$, then $d^{(2)}(\gamma, \eta) = 0$.
- V-statistic estimator of $d^{(2)}(\gamma, \eta)$:

$$\frac{1}{N^2}\mathrm{Tr}\left[(\mathbf{K}_x\mathbf{J}_N)^2\right] + \frac{1}{M^2}\mathrm{Tr}\left[(\mathbf{K}_y\mathbf{J}_M)^2\right] - \frac{2}{NM}\mathrm{Tr}\left[\mathbf{K}_{xy}\mathbf{J}_M\mathbf{K}_{xy}^\top\mathbf{J}_N\right],$$

with $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \gamma$, $(y_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \eta$, $\mathbf{K}_x = [k(x_i, x_j)]_{i,j=1}^N$, $\mathbf{K}_y = [k(y_i, y_j)]_{i,j=1}^M$, $\mathbf{K}_{x,y} = [k(x_i, y_j)]_{i,j=1}^{N,M}$, $\mathbf{J}_n = \mathbf{I}_n - \mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

Time complexity

Quadratic as MMD.

- $d^{(3)}(\gamma, \eta)$: similarly; quadratic time.

Cross-skewness independence criterion (CSIC)

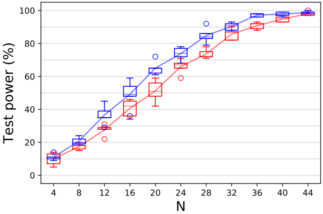
- By our theorem: if $\gamma = \gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2}$, then $\kappa^{(2,1)}_{k,\ell}(\gamma) = 0$ and $\kappa^{(1,2)}_{k,\ell}(\gamma) = 0$.
- V-statistic estimator of $\|\kappa^{(1,2)}_{k,\ell}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1} \otimes \mathcal{H}_\ell^{\otimes 2}}$:

$$\frac{1}{N^2}\Bigg\langle \mathbf{K} \circ \mathbf{K} \circ \mathbf{L} - 4\mathbf{K} \circ \mathbf{KH} \circ \mathbf{L} - 2\mathbf{K} \circ \mathbf{K} \circ \mathbf{LH} + 4\mathbf{KH} \circ \mathbf{K} \circ \mathbf{LH}$$
$$+2\mathbf{K} \circ \mathbf{L}\left\langle \frac{\mathbf{K}}{N^2}\right\rangle + 2\mathbf{KH} \circ \mathbf{HK} \circ \mathbf{L} + 4\mathbf{K} \circ \mathbf{HK} \circ \mathbf{LH} + \mathbf{K} \circ \mathbf{K}\left\langle \frac{\mathbf{L}}{N^2}\right\rangle$$
$$-8\mathbf{K} \circ \mathbf{LH}\left\langle \frac{\mathbf{K}}{N^2}\right\rangle - 4\mathbf{K} \circ \mathbf{HK}\left\langle \frac{\mathbf{L}}{N^2}\right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2}\right\rangle^2 \mathbf{L}\Bigg\rangle,$$

  with kernels $k : \mathcal{X}_1^2 \to \mathbb{R}$, $\ell : \mathcal{X}_2^2 \to \mathbb{R}$, $\mathbf{K} := \mathbf{K}_x, \mathbf{L} := \mathbf{L}_y$, $\langle \mathbf{A} \rangle := \sum_{i,j} A_{i,j}$.
- Time complexity: quadratic.

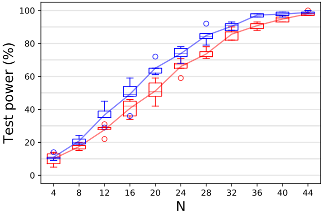
Numerical illustration: improved power

- Seoul bicycle rental data:
  - two-sample test (MMD, $d^{(2)}$): winter vs fall, $d = 11$,

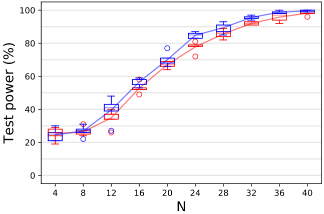


Numerical illustration: improved power

- Seoul bicycle rental data:
  - two-sample test (MMD, $d^{(2)}$): winter vs fall, $d = 11$,

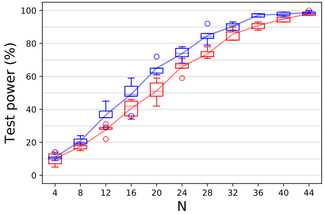


- Brazilian traffic data:
  - independence test (HSIC, CSIC); (blockage, fire, ...) vs slowness of traffic; $d_1 = 16$, $d_2 = 1$; l.h.s.

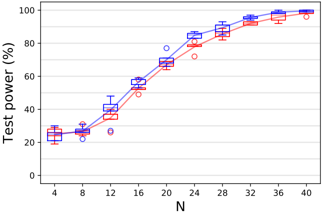


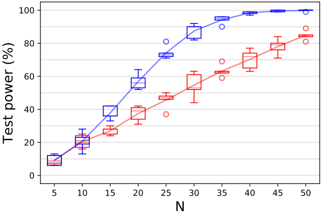
Numerical illustration: improved power

- Seoul bicycle rental data:
  - two-sample test (MMD, $d^{(2)}$): winter vs fall, $d = 11$,



- Brazilian traffic data:
  - independence test (HSIC, CSIC); (blockage, fire, ...) vs slowness of traffic; $d_1 = 16$, $d_2 = 1$; l.h.s.,
  - two-sample test (MMD, $d^{(3)}$): slow vs fast moving traffic, $d = 16$; r.h.s.





Summary

- We proposed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).

Summary

- We proposed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- MMD $\xleftarrow{\mathrm{m=d=1}}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d=2$).
- k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.

Summary

- We proposed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- MMD $\xleftarrow{\mathrm{m=d=1}}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d = 2$).
- k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.
- Relaxed kernel assumptions: point-separating.
- Higher-order cumulants: potential to improve power.
- TR on arXiv (submitted to NeurIPS), code.

Summary

- We proposed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- MMD $\xleftarrow{\mathrm{m=d=1}}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d = 2$).
- k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.
- Relaxed kernel assumptions: point-separating.
- Higher-order cumulants: potential to improve power.
- TR on arXiv (submitted to NeurIPS), code.

Appendix

- Bell numbers.
- Characteristic kernels.
- Universal kernels.
- Moments and cumulants on $\mathbb{R}^d$.
- Estimator for $d^{(3)}(\gamma, \eta)$.

Bell numbers

- $B(m) :=$ number of elements in $P(m)$.
- $B_0 = B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203,$ $B_7 = 877,\ B_8 = 4140,\ \ldots$

Bell numbers

- $B(m) :=$ number of elements in $P(m)$.
- $B_0 = B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203,$ $B_7 = 877,\ B_8 = 4140,\ \ldots$
- Recursion:

$$B_{m+1} = |P(m+1)| = \sum_{k=0}^{m} \binom{m}{k} B_k.$$

Bell numbers – continued

- Easy computation by the Bell triangle

$$
\begin{array}{lllll}
1 & & & & \\
1 & 2 & & & \\
2 & 3 & 5 & & \\
5 & 7 & 10 & 15 & \\
15 & 20 & 27 & 37 & 52 \\
52 & \dots & & &
\end{array}
$$

Bell numbers – continued

- Easy computation by the Bell triangle

$$
\begin{array}{ccccc}
1 & & & & \\
1 & 2 & & & \\
2 & 3 & 5 & & \\
5 & 7 & 10 & 15 & \\
15 & 20 & 27 & 37 & 52 \\
52 & \dots & & &
\end{array}
$$

- Asymptotics:

$$\frac{\ln B_n}{n} = \ln n - \ln\ln n - 1 + \frac{\ln\ln n}{\ln n} + \frac{1}{\ln n} + \frac{1}{2}\left(\frac{\ln\ln n}{\ln n}\right)^2 + \mathcal{O}\left(\frac{\ln\ln n}{\ln^2 n}\right)$$

as $n \to \infty$.

Contents

Description of characteristic kernels on $\mathbb{R}^d$

For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}', \boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega})$$

$(*)$: Bochner's theorem.

Description of characteristic kernels on $\mathbb{R}^d$

For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x},\mathbf{x}') = k_0(\mathbf{x}-\mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}',\boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}) \Rightarrow$$

$$\|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

$(*)$: Bochner's theorem, $c_{\mathbb{P}}$: characteristic function of $\mathbb{P}$.

Description of characteristic kernels on $\mathbb{R}^d$

For continuous bounded shift-invariant kernels on $\mathbb{R}^d$:

$$k(\mathbf{x},\mathbf{x}') = k_0(\mathbf{x}-\mathbf{x}') \overset{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{x}',\boldsymbol{\omega}\rangle} \mathrm{d}\Lambda(\boldsymbol{\omega}) \Rightarrow$$

$$\|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

$(*)$: Bochner's theorem, $c_{\mathbb{P}}$: characteristic function of $\mathbb{P}$.

Theorem ([Sriperumbudur et al., 2010])

k is characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0,1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0,1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

For $x \in \mathbb{R}^d$: $k_0(x)=\prod_{j=1}^d k_0(x_j)$, $\widehat{k_0}(\omega)=\prod_{j=1}^d \widehat{k_0}(\omega_j)$.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\Phi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{H}_k}$$

is a metric.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\Phi(x) = k(\cdot, x)$ is injective, i.e.

  $$\rho_k(x,y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{H}_k}$$

  is a metric.
- The normalized kernel (like corr)

  $$\tilde{k}(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

  is universal.

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^\infty \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r),\ r \in (0, \infty].$$

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^\infty \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r),\ r \in (0, \infty].$$

- If $a_n > 0\ \forall n$, then

$$k(\mathbf{x}, \mathbf{y}) = f(\langle \mathbf{x}, \mathbf{y} \rangle)$$

is universal on $\mathcal{X} := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \sqrt{r} \right\}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha \langle \mathbf{x}, \mathbf{y} \rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha\langle \mathbf{x}, \mathbf{y}\rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.
- $k(\mathbf{x}, \mathbf{y}) = e^{-\alpha\|\mathbf{x}-\mathbf{y}\|_2^2}$: exp. kernel & normalization.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = (1 - \langle \mathbf{x}, \mathbf{y} \rangle)^{-\alpha}$ binomial kernel
  - on $\mathcal{X}$ compact $\subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < 1\}$.
  - $f(t) = (1 - t)^{-\alpha} = \sum_{n=0}^{\infty} \underbrace{\binom{-\alpha}{n} (-1)^n}_{>0} t^n \quad (|t| < 1)$,

where $\binom{b}{n} = \sum_{i=1}^{n} \frac{b-i+1}{i}$.

Contents

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$
m-th moment	$\mu^{(m)}(\gamma)$	$\mu^m(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\deg(\mathbf{i})=m}$

where $\deg(\mathbf{i}) := i_1 + \cdots + i_d$, $\mu^0(\gamma) = 1$

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1}\cdots X_d^{i_d}\right] \in \mathbb{R}$
m-th moment	$\mu^{(m)}(\gamma)$	$\mu^m(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\deg(\mathbf{i})=m}$

and cumulants $\kappa(\gamma) = (\kappa^{\mathbf{i}}(\gamma))_{\mathbf{i}\in\mathbb{N}^d}$

$$\sum_{\mathbf{i}\in\mathbb{N}^d} \kappa^{\mathbf{i}}(\gamma)\frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!} = \log\left(\sum_{\mathbf{i}\in\mathbb{N}^d} \mu^{\mathbf{i}}(\gamma)\frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!}\right), \quad \boldsymbol{\theta} \in \mathbb{R}^d,$$

where $\deg(\mathbf{i}) := i_1 + \cdots + i_d$, $\mu^0(\gamma) = 1$, $\mathbf{i}! = i_1!\cdots i_d!$, $\boldsymbol{\theta}^{\mathbf{i}} = \theta_1^{i_1}\cdots\theta_d^{i_d}$.

Contents, moments and cumulants on $\mathbb{R}$, motivation of cumulants

Estimator for $d^{(3)}(\gamma, \eta) = \|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, $N = M$

$$d^{(3)}(\gamma, \eta) = \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} - 2\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}$$

Estimator for $d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, $N = M$

$$d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} - 2\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}$$

$$\begin{aligned}\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}} \approx \frac{1}{N^2}\Bigg\langle & \mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy} - 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\ & - 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} + 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{H}\mathbf{K}_{xy} \\ & + 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\left\langle \frac{\mathbf{K}_{xy}}{N^2}\right\rangle + 2\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\ & + 2\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{K}_{xy}\mathbf{H} - 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H}\left\langle \frac{\mathbf{K}_{xy}}{N^2}\right\rangle \\ & - 6\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy}\left\langle \frac{\mathbf{K}_{xy}}{N^2}\right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2}\right\rangle^2 \mathbf{K}_{xy}\Bigg\rangle.\end{aligned}$$

Note: Matrix multiplication takes precedence over the Hadamard one.

Estimator for $d^{(3)}(\gamma,\eta)$ – continued

$$\|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} \approx \frac{1}{N^2}\Big\langle \mathbf{K}_x \circ \mathbf{K}_x \circ \mathbf{K}_x - 6\mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x$$
$$+ 4\mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} + 3\mathbf{K}_x \circ \mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle$$
$$+ 6\mathbf{K}_x\mathbf{H} \circ \mathbf{H}\mathbf{K}_x \circ \mathbf{K}_x - 12\mathbf{K}_x \circ \mathbf{H}\mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle$$
$$+ 4\left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle^2 \mathbf{K}_x \Big\rangle.$$

$\|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$: similarly (change $\mathbf{K}_x$ to $\mathbf{K}_y$).

Contents, $d^2(\gamma,\eta)$ estimation

Billingsley, P. (2012).
Probability and Measure.
Wiley.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 6(3):67–93.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.