Online Change Detection via Random Fourier Features

Shakeel Gavioli-Akilagun

London School of Economics
Department of Statistics

Collaborator

Florian Kalinke @
Karlsruher Institut für Technologie

Problem Statement

- **Setup:** the sequence $X_1, X_2, \ldots$ is observed online. The X's are defined on $\mathbb{R}^d$, and $\exists \eta \in \mathbb{N}$ (possibly infinite) and two measures $\mathbb{P}, \mathbb{Q} \in M_1^+\left(\mathbb{R}^d\right)$ for which

$$X_t \overset{\text{i.i.d.}}{\sim} \begin{cases} \mathbb{P} & \text{for } t = 1, \ldots, \eta \\ \mathbb{Q} & \text{for } t = \eta + 1, \eta + 2, \ldots \end{cases}.$$

- **Goal:** stop the process with minimal delay as soon as η is reached, but not before.

Problem Statement

- **Setup:** the sequence $X_1, X_2, \ldots$ is observed online. The X's are defined on $\mathbb{R}^d$, and $\exists \eta \in \mathbb{N}$ (possibly infinite) and two measures $\mathbb{P}, \mathbb{Q} \in M_1^+\left(\mathbb{R}^d\right)$ for which

$$X_t \overset{\text{i.i.d.}}{\sim} \begin{cases} \mathbb{P} & \text{for } t = 1, \ldots, \eta \\ \mathbb{Q} & \text{for } t = \eta + 1, \eta + 2, \ldots \end{cases}.$$

- **Goal:** stop the process with minimal delay as soon as η is reached, but not before.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Example: spot the parcel theft

- **Goal:** alert the user when unusual activity is detected in front of their house.

Problem Statement (continued)

- **Formally:** the aim is to test the following hypotheses in an online fashion

$$H_{0,n} : X_t \sim \mathbb{P} \text{ for each } t \leqslant n \text{ and some } \mathbb{P} \in M_1^+\left(\mathbb{R}^d\right)$$

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+(\mathbb{R})\,.$$

- **Goal:** construct an extended stopping time N which is guaranteed to be "close" to η and which satisfies either of
  1. Average run length: $\mathbb{E}_\infty[N] \geqslant \gamma$ for some $\gamma > 1$,
  2. Uniform false alarm rate: $\mathbb{P}_\infty(N \leqslant \infty) \leqslant \alpha$ for some $\alpha \in (0,1)$.

Problem Statement (continued)

- **Formally:** the aim is to test the following hypotheses in an online fashion

$$H_{0,n} : X_t \sim \mathbb{P} \text{ for each } t \leqslant n \text{ and some } \mathbb{P} \in M_1^+\left(\mathbb{R}^d\right)$$

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+ (\mathbb{R}) .$$

- **Goal:** construct an extended stopping time N which is guaranteed to be "close" to η and which satisfies either of
  1. Average run length: $\mathbb{E}_\infty [N] \geqslant \gamma$ for some $\gamma > 1$,
  2. Uniform false alarm rate: $\mathbb{P}_\infty (N \leqslant \infty) \leqslant \alpha$ for some $\alpha \in (0, 1)$.

Problem Statement (continued)

- **Formally:** the aim is to test the following hypotheses in an online fashion

$$H_{0,n} : X_t \sim \mathbb{P} \text{ for each } t \leqslant n \text{ and some } \mathbb{P} \in M_1^+\left(\mathbb{R}^d\right)$$

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+(\mathbb{R}).$$

- **Goal:** construct an extended stopping time N which is guaranteed to be "close" to η and which satisfies either of
  1. Average run length: $\mathbb{E}_\infty[N] \geqslant \gamma$ for some $\gamma > 1$,
  2. Uniform false alarm rate: $\mathbb{P}_\infty(N \leqslant \infty) \leqslant \alpha$ for some $\alpha \in (0,1)$.

Problem Statement (continued)

- **Formally:** the aim is to test the following hypotheses in an online fashion

$$H_{0,n} : X_t \sim \mathbb{P} \text{ for each } t \leqslant n \text{ and some } \mathbb{P} \in M_1^+\left(\mathbb{R}^d\right)$$

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+\left(\mathbb{R}\right).$$

- **Goal:** construct an extended stopping time N which is guaranteed to be "close" to η and which satisfies either of
  1. Average run length: $\mathbb{E}_\infty\left[N\right] \geqslant \gamma$ for some $\gamma > 1$,
  2. Uniform false alarm rate: $\mathbb{P}_\infty\left(N \leqslant \infty\right) \leqslant \alpha$ for some $\alpha \in (0, 1)$.

Reproducing Kernel Hilbert Spaces (RKHSs)

Definition (RKHSs)

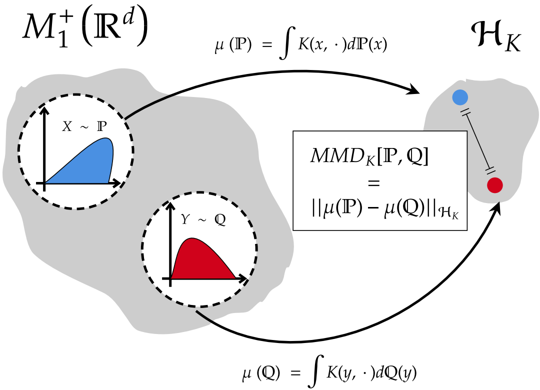
A Hilbert space $\mathcal{H}_K$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and norm $\|\cdot\|_{\mathcal{H}_K}$ consisting of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ is called an RKHS if there exists a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ for which

- $K(x, \cdot) \in \mathcal{H}_K$ for all $x \in \mathbb{R}^d$
- $f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}_K}$ for all $x \in \mathbb{R}^d$ and all $f \in \mathcal{H}_K$ (reproducing property).

- **Kernel methods:** represent data as elements of $\mathcal{H}_K$ using $K(x, \cdot)$, do learning in $\mathcal{H}_K$. Due to the reproducing everything can be expressed in terms of $K(x, y) \Rightarrow$ actually computable.

Maximum Mean Discrepancy

- The Maximum Mean Discrepancy (Gretton u. a., 2012, MMD) measures discrepancies between distributions by considering their distance in RKHS norm.



Characteristic Kernels

- If $\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}] = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ then $\mathcal{H}_K$ is called characteristic. Sriperumbudur u. a. (2010) have shown that $\mathcal{H}_K$ is characteristic if
  - (C1) $\sup_x \sqrt{K(x,x)} \leqslant C$ for some $C > 0$ (bounded)
  - (C2) $K(x,y) = \psi(x-y)$ for some positive definite ψ (translation invariant)
  - (C3) $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(x) = \int e^{-i\omega' x} \mathrm{d}\Lambda(\omega)$ (spectrum support)
- Some examples of characteristic kernels include...

Kernel	$\psi(x)$	$\Lambda(\omega)$	supp(Λ)
Gaussian	$e^{-x^2/(2\sigma^2)}$	$\sigma e^{-\sigma^2\omega^2/2}$	$\mathbb{R}$
Laplace	$e^{-\sigma\|x\|}$	$\frac{2}{\pi}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}−spline	$*_1^{(2n+1)} 1_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$	$\mathbb{R}$

... to extend to $\mathbb{R}^d$ take products.

Characteristic Kernels

- If $\mathrm{MMD}_K[\mathbb{P},\mathbb{Q}] = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ then $\mathcal{H}_K$ is called characteristic. Sriperumbudur u. a. (2010) have shown that $\mathcal{H}_K$ is characteristic if
  (C1) $\sup_x \sqrt{K(x,x)} \leqslant C$ for some $C > 0$ (bounded)
  (C2) $K(x,y) = \psi(x-y)$ for some positive definite ψ (translation invariant)
  (C3) $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(x) = \int e^{-i\omega' x} \mathrm{d}\Lambda(\omega)$ (spectrum support)
- Some examples of characteristic kernels include...

Kernel	$\psi(x)$	$\Lambda(\omega)$	$\mathrm{supp}(\Lambda)$
Gaussian	$e^{-x^2/(2\sigma^2)}$	$\sigma e^{-\sigma^2\omega^2/2}$	$\mathbb{R}$
Laplace	$e^{-\sigma\lvert x\rvert}$	$\frac{2}{\pi}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}−spline	$*_1^{(2n+1)} 1_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$	$\mathbb{R}$

... to extend to $\mathbb{R}^d$ take products.

Characteristic Kernels

- If $\mathrm{MMD}_K[\mathbb{P},\mathbb{Q}] = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ then $\mathcal{H}_K$ is called characteristic. Sriperumbudur u. a. (2010) have shown that $\mathcal{H}_K$ is characteristic if
  (C1) $\sup_x \sqrt{K(x,x)} \leqslant C$ for some $C > 0$ (bounded)
  (C2) $K(x,y) = \psi(x-y)$ for some positive definite ψ (translation invariant)
  (C3) $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(x) = \int e^{-i\omega' x} \mathrm{d}\Lambda(\omega)$ (spectrum support)
- Some examples of characteristic kernels include...

Kernel	$\psi(x)$	$\Lambda(\omega)$	$\mathrm{supp}(\Lambda)$
Gaussian	$e^{-x^2/(2\sigma^2)}$	$\sigma e^{-\sigma^2\omega^2/2}$	$\mathbb{R}$
Laplace	$e^{-\sigma\|x\|}$	$\frac{2}{\pi}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}–spline	$*_1^{(2n+1)} 1_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$	$\mathbb{R}$

... to extend to $\mathbb{R}^d$ take products.

Characteristic Kernels

- If $\mathrm{MMD}_K[\mathbb{P},\mathbb{Q}] = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ then $\mathcal{H}_K$ is called characteristic. Sriperumbudur u. a. (2010) have shown that $\mathcal{H}_K$ is characteristic if
  - (C1) $\sup_x \sqrt{K(x,x)} \leqslant C$ for some $C > 0$ (bounded)
  - (C2) $K(x,y) = \psi(x-y)$ for some positive definite ψ (translation invariant)
  - (C3) $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(x) = \int e^{-i\omega' x} \mathrm{d}\Lambda(\omega)$ (spectrum support)
- Some examples of characteristic kernels include...

Kernel	$\psi(x)$	$\Lambda(\omega)$	$\mathrm{supp}(\Lambda)$
Gaussian	$e^{-x^2/(2\sigma^2)}$	$\sigma e^{-\sigma^2\omega^2/2}$	$\mathbb{R}$
Laplace	$e^{-\sigma\|x\|}$	$\frac{2}{\pi}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}−spline	$*_1^{(2n+1)} 1_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$	$\mathbb{R}$

... to extend to $\mathbb{R}^d$ take products.

Characteristic Kernels

- If $\text{MMD}_K[\mathbb{P},\mathbb{Q}] = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ then $\mathcal{H}_K$ is called characteristic. Sriperumbudur u. a. (2010) have shown that $\mathcal{H}_K$ is characteristic if
  (C1) $\sup_x \sqrt{K(x,x)} \leqslant C$ for some $C > 0$ (bounded)
  (C2) $K(x,y) = \psi(x-y)$ for some positive definite ψ (translation invariant)
  (C3) $\text{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(x) = \int e^{-i\omega' x} \mathrm{d}\Lambda(\omega)$ (spectrum support)
- Some examples of characteristic kernels include...

Kernel	$\psi(x)$	$\Lambda(\omega)$	$\text{supp}(\Lambda)$
Gaussian	$e^{-x^2/(2\sigma^2)}$	$\sigma e^{-\sigma^2\omega^2/2}$	$\mathbb{R}$
Laplace	$e^{-\sigma\lvert x\rvert}$	$\frac{2}{\pi}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}−spline	$*_1^{(2n+1)} 1_{[-\frac{1}{2},\frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}(\omega/2)}{\omega^{2n+2}}$	$\mathbb{R}$

... to extend to $\mathbb{R}^d$ take products.

Maximum Mean Discrepancy (continued)

- Given independent samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$ a natural estimator of the (squared) MMD is given by

$$\mathrm{MMD}_K^2\left[X_{1:n}, Y_{1:m}\right] = \left\| \mu(\hat{\mathbb{P}}_{1:n}) - \mu(\hat{\mathbb{Q}}_{1:m}) \right\|_{\mathcal{H}}^2 .$$

- Computing the above requires $\mathcal{O}\left(n^2 + m^2\right)$ basic operations making its use impractical for online problems.

Maximum Mean Discrepancy (continued)

- Given independent samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$ a natural estimator of the (squared) MMD is given by

$$\mathrm{MMD}_K^2 \left[X_{1:n}, Y_{1:m}\right] = \left\| \frac{1}{n} \sum_{i=1}^{n} K\left(X_i, \cdot\right) - \frac{1}{m} \sum_{j=1}^{m} K\left(Y_j, \cdot\right) \right\|_{\mathcal{H}}^2 .$$

- Computing the above requires $\mathcal{O}\left(n^2 + m^2\right)$ basic operations making its use impractical for online problems.

Maximum Mean Discrepancy (continued)

- Given independent samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$ a natural estimator of the (squared) MMD is given by

$$\mathrm{MMD}_K^2 \left[X_{1:n}, Y_{1:m}\right] = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(X_i, X_j\right) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} K\left(Y_i, Y_j\right) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} K\left(X_i, Y_j\right).$$

- Computing the above requires $\mathcal{O}\left(n^2 + m^2\right)$ basic operations making its use impractical for online problems.

Random Fourier Features

- Let K satisfy $C1$-$C3$. By Bochner's theorem

$$K(x, y) = \int_{\mathbb{R}^d} e^{i\omega'(x-y)} \mathrm{d}\Lambda(\omega).$$

- Rahimi und Recht (2007) note that as both Λ, K are real the integrand must be real and by Euler's identity can be replaced with $\cos(\omega'(x-y))$. Thus

$$K(x, y) = \int_{\mathbb{R}^d} \cos(\omega'(x-y)) \mathrm{d}\Lambda(\omega) = \mathbb{E}_{\omega\sim\Lambda}\left[\cos(\omega'(x-y))\right].$$

- Then for $r \in \mathbb{N}$ and $\omega_1, \ldots, \omega_r \overset{i.i.d}{\sim} \Lambda$ an unbiased estimator for $K(x, y)$ is

$$\hat{K}(x, y) = \frac{1}{r} \sum_{k=1}^{r} \cos(\omega_k'(x-y)).$$

Random Fourier Features

- Let K satisfy $C1$-$C3$. By Bochner's theorem

$$K(x,y) = \int_{\mathbb{R}^d} e^{i\omega'(x-y)} \mathrm{d}\Lambda(\omega).$$

- Rahimi und Recht (2007) note that as both Λ, K are real the integrand must be real and by Euler's identity can be replaced with $\cos(\omega'(x-y))$. Thus

$$K(x,y) = \int_{\mathbb{R}^d} \cos\left(\omega'(x-y)\right) \mathrm{d}\Lambda(\omega) = \mathbb{E}_{\omega\sim\Lambda}\left[\cos\left(\omega'(x-y)\right)\right].$$

- Then for $r \in \mathbb{N}$ and $\omega_1, \ldots, \omega_r \overset{i.i.d}{\sim} \Lambda$ an unbiased estimator for $K(x,y)$ is

$$\hat{K}(x,y) = \frac{1}{r}\sum_{k=1}^{r} \cos\left(\omega_k'(x-y)\right).$$

Random Fourier Features

- Let K satisfy $C1$-$C3$. By Bochner's theorem

$$K\left(x,y\right)=\int_{\mathbb{R}^d}e^{i\omega'(x-y)}\mathrm{d}\Lambda\left(\omega\right).$$

- Rahimi und Recht (2007) note that as both Λ, K are real the integrand must be real and by Euler's identity can be replaced with $\cos\left(\omega'\left(x-y\right)\right)$. Thus

$$K\left(x,y\right)=\int_{\mathbb{R}^d}\cos\left(\omega'(x-y)\right)\mathrm{d}\Lambda\left(\omega\right)=\mathbb{E}_{\omega\sim\Lambda}\left[\cos\left(\omega'(x-y)\right)\right].$$

- Then for $r\in\mathbb{N}$ and $\omega_1,\ldots,\omega_r \overset{i.i.d}{\sim} \Lambda$ an unbiased estimator for $K(x,y)$ is

$$\hat{K}\left(x,y\right)=\frac{1}{r}\sum_{k=1}^{r}\cos\left(\omega_k'(x-y)\right).$$

Random Fourier Features

- For $r \in \mathbb{N}$ and $\omega_1, \dots, \omega_r \overset{i.i.d}{\sim} \Lambda$ given independent samples $\{X_1, \dots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \dots, Y_m\} \sim \mathbb{Q}$ a simple unbiased estimator of $\mathrm{MMD}^2_K[X_{1:n}, Y_{1:m}]$ is

$$\mathrm{MMD}^2_{\hat{K}}[X_{1:n}, Y_{1:m}] = \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{1}{r}\sum_{k=1}^{r} z_{\omega_k}(X_i) - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{r}\sum_{k=1}^{r} z_{\omega_k}(Y_j) \right\|_2^2 .$$

  where $z_\omega(x) = (\cos(\omega' x), \sin(\omega' x))'$.
- Importantly, this quantity can be computed in $\mathcal{O}(rn + rm)$ basic operations, and updated in $\mathcal{O}(r)$ time, making it ideal for online problems.

Random Fourier Features

- For $r \in \mathbb{N}$ and $\omega_1, \ldots, \omega_r \overset{i.i.d}{\sim} \Lambda$ given independent samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$ a simple unbiased estimator of $\mathrm{MMD}^2_K [X_{1:n}, Y_{1:m}]$ is

$$\mathrm{MMD}^2_{\hat{K}} [X_{1:n}, Y_{1:m}] = \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r} \sum_{k=1}^{r} z_{\omega_k}(X_i) - \frac{1}{m} \sum_{j=1}^{m} \frac{1}{r} \sum_{k=1}^{r} z_{\omega_k}(Y_j) \right\|_2^2 .$$

  where $z_\omega(x) = (\cos(\omega' x), \sin(\omega' x))'$.
- Importantly, this quantity can be computed in $\mathcal{O}(rn + rm)$ basic operations, and updated in $\mathcal{O}(r)$ time, making it ideal for online problems.

Random Fourier Features

- For $r \in \mathbb{N}$ and $\omega_1, \ldots, \omega_r \overset{i.i.d}{\sim} \Lambda$ given independent samples $\{X_1, \ldots, X_n\} \sim \mathbb{P}$ and $\{Y_1, \ldots, Y_m\} \sim \mathbb{Q}$ a simple unbiased estimator of $\mathrm{MMD}^2_K\left[X_{1:n}, Y_{1:m}\right]$ is

$$\mathrm{MMD}^2_{\hat{K}}\left[X_{1:n}, Y_{1:m}\right] = \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{K}(X_i, \cdot) - \frac{1}{m} \sum_{j=1}^{m} \hat{K}(Y_j, \cdot) \right\|_2^2 .$$

- Importantly, this quantity can be computed in $\mathcal{O}\left(rn + rm\right)$ basic operations, and updated in $\mathcal{O}(r)$ time, making it ideal for online problems.

Online Change Detection via Random Fourier Features

- Recall, the aim is to test H_0 in an online manner against

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+(\mathbb{R}).$$

- This can be achieved with the statistic

$$\max_{\tau=1\dots n-1} \sqrt{\frac{\tau(n-\tau)}{n}} \mathrm{MMD}_{\hat{K}} \left[X_{1:\tau}, X_{(\tau+1):n} \right].$$

- We use a dyadic approximation scheme due to Lai (1995)

$$N = \inf \left\{ n \geqslant 2 \mid \bigvee_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \sqrt{\frac{2^j(n-2^j)}{n}} \mathrm{MMD}_{\hat{K}} \left[X_{1:(n-2^j)}, X_{(n-2^j+1):n} \right] > \lambda_n \right\}.$$

Online Change Detection via Random Fourier Features

- Recall, the aim is to test H_0 in an online manner against

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+(\mathbb{R}).$$

- This can be achieved with the statistic

$$\max_{\tau = 1 \dots n-1} \sqrt{\frac{\tau(n-\tau)}{n}} \mathrm{MMD}_{\hat{K}} \left[X_{1:\tau}, X_{(\tau+1):n} \right].$$

- We use a dyadic approximation scheme due to Lai (1995)

$$N = \inf \left\{ n \geqslant 2 \mid \bigvee_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \sqrt{\frac{2^j(n-2^j)}{n}} \mathrm{MMD}_{\hat{K}} \left[X_{1:(n-2^j)}, X_{(n-2^j+1):n} \right] > \lambda_n \right\}.$$

Online Change Detection via Random Fourier Features

- Recall, the aim is to test H_0 in an online manner against

$$H_{1,n} : \exists \eta < n \text{ s.t } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leqslant t \leqslant \eta \\ \mathbb{Q} & \text{if } \eta < t \leqslant n \end{cases}, \text{ and } \mathbb{P}, \mathbb{Q} \in M_1^+(\mathbb{R}).$$

- This can be achieved with the statistic

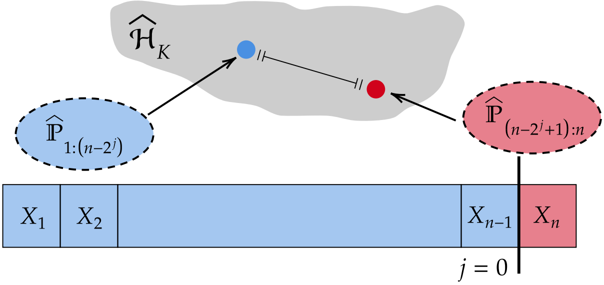
$$\max_{\tau=1\ldots n-1} \sqrt{\frac{\tau(n-\tau)}{n}} \mathrm{MMD}_{\widehat{K}}\left[X_{1:\tau}, X_{(\tau+1):n}\right].$$

- We use a dyadic approximation scheme due to Lai (1995)

$$N = \inf\left\{ n \geqslant 2 \mid \bigvee_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \sqrt{\frac{2^j(n-2^j)}{n}} \mathrm{MMD}_{\widehat{K}}\left[X_{1:(n-2^j)}, X_{(n-2^j+1):n}\right] > \lambda_n \right\}.$$

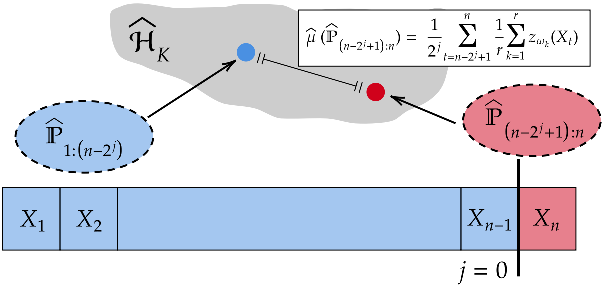
Online Change Detection via Random Fourier Features

- Having observed data $\{X_1, \ldots, X_n\}$, we consider $\log_2(n)$ possible sample splits. For every such split we approximate the MMD between empirical measures of the two samples using RFFs, and stop the process at the first n for which at least one such statistic is larger than a given threshold.



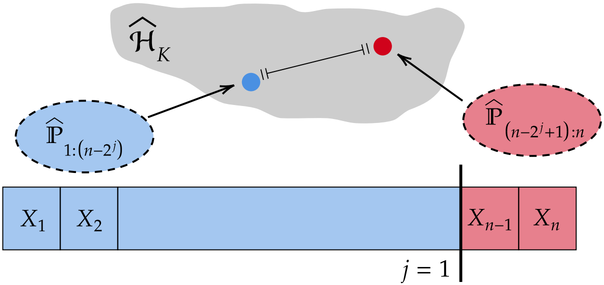
Online Change Detection via Random Fourier Features

- Having observed data $\{X_1, \ldots, X_n\}$, we consider $\log_2(n)$ possible sample splits. For every such split we approximate the MMD between empirical measures of the two samples using RFFs, and stop the process at the first n for which at least one such statistic is larger than a given threshold.



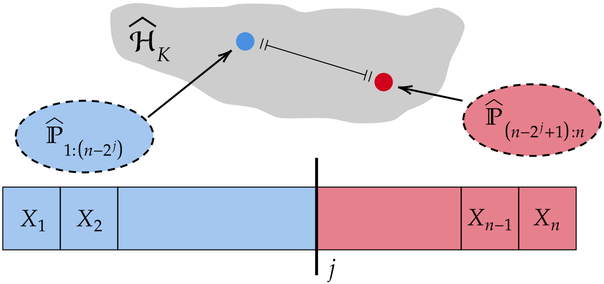
Online Change Detection via Random Fourier Features

- Having observed data $\{X_1, \ldots, X_n\}$, we consider $\log_2(n)$ possible sample splits. For every such split we approximate the MMD between empirical measures of the two samples using RFFs, and stop the process at the first n for which at least one such statistic is larger than a given threshold.



Online Change Detection via Random Fourier Features

- Having observed data $\{X_1, \ldots, X_n\}$, we consider $\log_2(n)$ possible sample splits. For every such split we approximate the MMD between empirical measures of the two samples using RFFs, and stop the process at the first n for which at least one such statistic is larger than a given threshold.



Explicit Algorithm

Algorithm 1: RFF MMD Change Detector

Data: $X_1, X_2, \ldots, \alpha \in (0,1)$

Result: Changepoint location and detection time

$\mathcal{W} \leftarrow$ empty list;

for $X_t \in X_1, X_2, \ldots$;   `/* Main loop */`

do

     $W.c \leftarrow 1$;

     $W.z \leftarrow z_k(X_t)$;

     $\mathcal{W} \leftarrow \mathcal{W}.\text{append}(W)$;

     for $i \in 1, \ldots, |\mathcal{W}| - 1$;   `/* Detect changes */`

     do

         $n \leftarrow \sum_{j=i+1}^{|\mathcal{W}|} \mathcal{W}_j.c$;

         $m \leftarrow \sum_{j=1}^{i} \mathcal{W}_j.c$;

         $\text{MMD}_{\hat{k}} \leftarrow \left\| \frac{1}{n} \sum_{j=i+1}^{|\mathcal{W}|} \mathcal{W}_j.z - \frac{1}{m} \sum_{j=1}^{i} \mathcal{W}_j.z \right\|_2$;

         $\alpha' \leftarrow \alpha / (|\mathcal{W}| - 1)$;

         if $\sqrt{\frac{nm}{n+m}} \text{MMD}_{\hat{K}} \geq \lambda$ **then**

             print *Change detected at element* X_t; *most likely at* i ;

             Drop tail of $\mathcal{W}$;

         end

     end

     while $|\mathcal{W}| \geq 2$;   `/* Maintain exponential structure */`

     do

         $W_1 \leftarrow$ pop $\mathcal{W}$;

         $W_2 \leftarrow$ pop $\mathcal{W}$;

         if $W_1.c = W_2.c$ **then**

             $W.c \leftarrow W_1.c + W_2.c$;

             $W.z \leftarrow W_1.z + W_2.z$;

             $\mathcal{W} \leftarrow \mathcal{W}.\text{append}(W)$;

         else

             break ;

     end

end

Time Complexity: $\mathcal{O}(r\log(n))$ *per iteration*

- **Setup:** The computation of $z_{\omega_k}(X)$ requires computing $2r$ trigonometric functions of d-dimensional inner products and thus is in $\mathcal{O}(rd)$.
- **Change detection:** The memoization of all sums allows to implement the change detection in a single sweep over $\mathcal{W}$; at each step, the attributes of one $W \in \mathcal{W}$ are subtracted from one sum and added to another sum. This gives a total complexity of $\mathcal{O}(r\log(n))$
- **Maintenance:** In the worst case, $\mathcal{O}(\log(n))$ merge operations need to be performed. Each merge requires $\mathcal{O}(r)$ operations, which yields a total cost of $\mathcal{O}(r\log(n))$.

Time Complexity: $\mathcal{O}(r \log(n))$ *per iteration*

- **Setup:** The computation of $z_{\omega_k}(X)$ requires computing $2r$ trigonometric functions of d-dimensional inner products and thus is in $\mathcal{O}(rd)$.
- **Change detection:** The memoization of all sums allows to implement the change detection in a single sweep over $\mathcal{W}$; at each step, the attributes of one $W \in \mathcal{W}$ are subtracted from one sum and added to another sum. This gives a total complexity of $\mathcal{O}(r \log(n))$
- **Maintenance:** In the worst case, $\mathcal{O}(\log(n))$ merge operations need to be performed. Each merge requires $\mathcal{O}(r)$ operations, which yields a total cost of $\mathcal{O}(r \log(n))$.

Time Complexity: $\mathcal{O}(r \log(n))$ *per iteration*

- **Setup:** The computation of $z_{\omega_k}(X)$ requires computing $2r$ trigonometric functions of d-dimensional inner products and thus is in $\mathcal{O}(rd)$.
- **Change detection:** The memoization of all sums allows to implement the change detection in a single sweep over $\mathcal{W}$; at each step, the attributes of one $W \in \mathcal{W}$ are subtracted from one sum and added to another sum. This gives a total complexity of $\mathcal{O}(r \log(n))$
- **Maintenance:** In the worst case, $\mathcal{O}(\log(n))$ merge operations need to be performed. Each merge requires $\mathcal{O}(r)$ operations, which yields a total cost of $\mathcal{O}(r \log(n))$.

Statistical Size

Theorem (average run length)

Let N be the extended stopping time defined previously. For any $\gamma > 1$, if the sequence of thresholds satisfies

$$\lambda_n \geqslant \sqrt{2} + \sqrt{2\log\left(4\gamma\log_2\left(2\gamma\right)\right)} \quad n \in \mathbb{N}$$

it holds that $\mathbb{E}_\infty\left[N\right] \geqslant \gamma$.

- Note that the above result does not depend on the number of random Fourier features used to approximate the MMD.

Statistical Size (continued)

Theorem (uniform false alarm rate)

Let N be the extended stopping time defined previously. For any $\alpha \in (0,1)$, if the sequence of thresholds satisfies

$$\lambda_n \geqslant \sqrt{2} + \sqrt{2\left(\log(n/\alpha) + 2\log\left(\log_2(n)\right) + \log\left(\log_2(2n)\right)\right)} \quad n \in \mathbb{N}$$

then it holds that $\mathbb{P}_\infty\left(N < \infty\right) \leqslant \alpha$.

- Note that the above result does not depend on the number of random Fourier features used to approximate the MMD.

Statistical Power

Theorem (high probability detection delay)

Let λ_n be as defined in the previous theorem. If supp$(\mathbb{P}) \cup$ supp$(\mathbb{Q}) \subseteq \mathcal{X}$ for some compact set $\mathcal{X} \subset \mathbb{R}^d$,

$$\eta \geqslant \frac{C_1 \log(2\eta/\alpha)}{MMD_K^2[\mathbb{P}, \mathbb{Q}]},$$

and moreover the number of random features is chosen so that

$$\sqrt{r} \geqslant \frac{C_2 + C_3\sqrt{2\log(2/\alpha)}}{MMD_K^2[\mathbb{P}, \mathbb{Q}]}$$

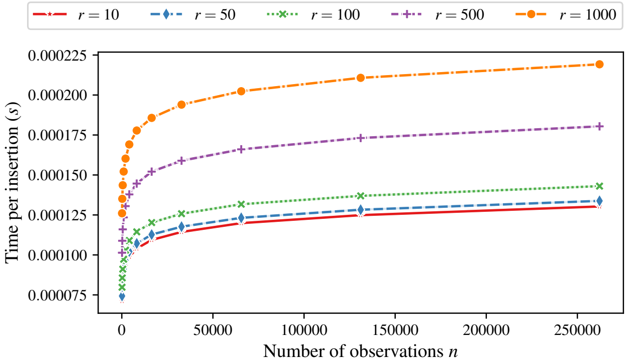
then with probability at least $1-\alpha$ it holds that

$$(N-\eta)^+ \leqslant 1 \vee \frac{C_4 \log(2\eta/\alpha)}{MMD_K^2[\mathbb{P}, \mathbb{Q}]}$$

where C_1, C_2, C_3, and C_4 are absolute constants.

Runtime Experiments

- Average runtime (10 repetitions) under the null of no change of the RFF-MMD algorithm using $r \in \{10, 50, \ldots, 1000\}$ random Fourier features in dimension $d = 1$.



Competing Methods

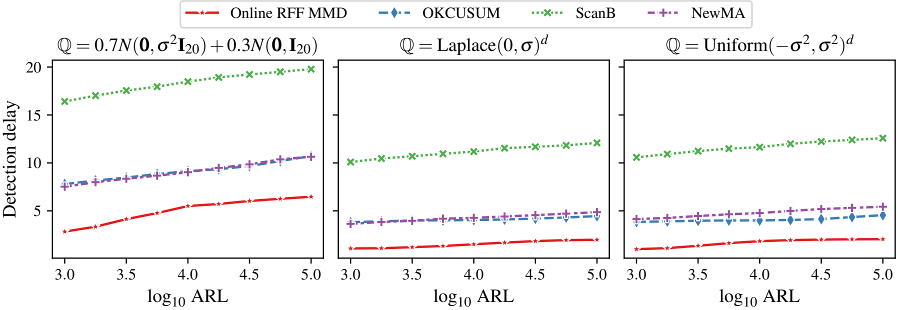
- We compare with three state of the are methods for online change detection...

Name	Time complexity	Approach	Training data
RFF-MMD	$r\log(n)$	RFF + dyadic scheme	No
ScanB	NW^2	sliding window	Yes
OKCUSUM	NW^2	max over multiple windows	Yes
NewMA	rd	RFF + exponential moving average	No

... for ScanB and OKCUSUM: N denotes the number of windows and W denotes the (max) window size.

Average Detection Delay Experiments

- Average detection delay (1000 repetitions) from 64 samples of $\mathbb{P} = \mathcal{N}(\mathbf{0}_{20}, \boldsymbol{I}_{20})$ to the distribution indicated on top. $r = 1000$ random Fourier features are used to approximate the MMD, and thresholds for each method are calibated vai Monte Carlo.



Thank you!

References I

[Gretton u. a. 2012] GRETTON, Arthur ; BORGWARDT, Karsten M. ; RASCH, Malte J. ; SCHÖLKOPF, Bernhard ; SMOLA, Alexander: A kernel two-sample test. In: *The Journal of Machine Learning Research* 13 (2012), Nr. 1, S. 723–773

[Lai 1995] LAI, Tze L.: Sequential changepoint detection in quality control and dynamical systems. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1995), Nr. 4, S. 613–644

[Rahimi und Recht 2007] RAHIMI, Ali ; RECHT, Benjamin: Random features for large-scale kernel machines. In: *Advances in neural information processing systems* 20 (2007)

[Sriperumbudur u. a. 2010] SRIPERUMBUDUR, Bharath K. ; GRETTON, Arthur ; FUKUMIZU, Kenji ; SCHÖLKOPF, Bernhard ; LANCKRIET, Gert R.: Hilbert space embeddings and metrics on probability measures. In: *The Journal of Machine Learning Research* 11 (2010), S. 1517–1561