Karlsruhe Institute of Technology

Optimal Online Change Detection via Random Fourier Features

Florian Kalinke [1] Shakeel Gavioli-Akilagun [2]

[1]Karlsruhe Institute of Technology [2]City University of Hong Kong

香港城市大學 City University of Hong Kong

Quick Summary

- New non-parametric online change detection algorithm for $\mathbb{R}^d$-valued data with runtime complexity of $O(\log n)$ per observation.
- Guarantees on average run length, uniform false alarm probability, and expected detection delay.
- Minimax optimality of expected detection delay.
- Experimental validation on synthetic, MNIST, HASC, and audio data.
- Key idea: Online approximation of the maximum mean discrepancy on a dyadic grid using random Fourier features.

Problem Statement

- **Setup:** $X_1, X_2, \dots; X_t \in \mathbb{R}^d$; $\mathbb{P}, \mathbb{Q}$ probability measures on $\mathbb{R}^d$; $\mathbb{P} \neq \mathbb{Q}$. $\exists \eta \in \mathbb{N} \cup \{\infty\}$ such that

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

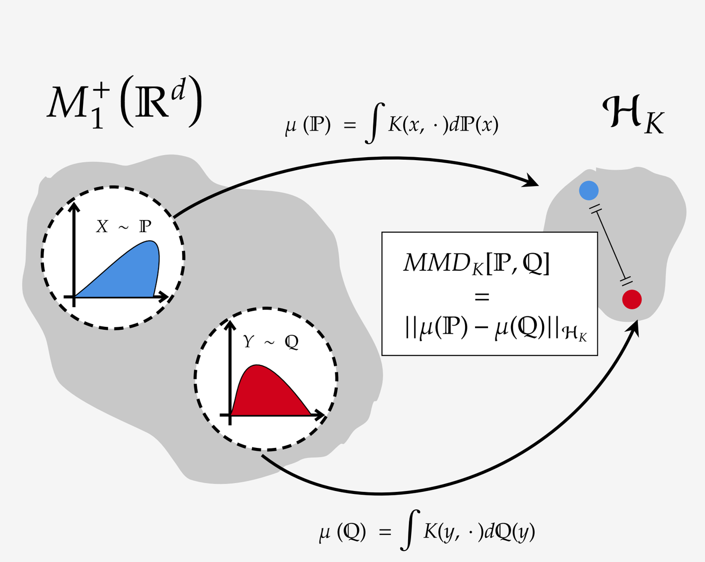
- **Goal:** Stop with minimal delay as soon as η is reached, but not before; never stop in case $\eta = \infty$.

Maximum Mean Discrepancy (MMD)

- Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a bounded kernel with associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ and (canonical) feature map $K(\cdot, \mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$. Then

$$\mathrm{MMD}_K(\mathbb{P}, \mathbb{Q}) = \|\mu_K(\mathbb{P}) - \mu_K(\mathbb{Q})\|_{\mathcal{H}_K},$$

where $\mu_K : \mathbb{P} \mapsto \int K(\cdot, \mathbf{x}) \mathrm{d}\mathbb{P}(\mathbf{x})$ is the kernel mean embedding [1].
- Kernel-based metric on probability measures under mild conditions.
- Classic estimators are $O(n^2)$.



Random Fourier Feature (RFF) Approximation

- If K is bounded continuous translation-invariant, by Bochner's theorem

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

- Approximate K by [3; 4]

$$\hat{K}(\mathbf{x}, \mathbf{y}) := \langle \hat{z}_K(\mathbf{x}), \hat{z}_K(\mathbf{y}) \rangle, \text{ where } \hat{z}_K(\mathbf{x}) = \frac{1}{\sqrt{r}} \left((\sin(\omega_j^\top \mathbf{x}), \cos(\omega_j^\top \mathbf{x})) \right)_{j=1}^r \in \mathbb{R}^{2r},$$

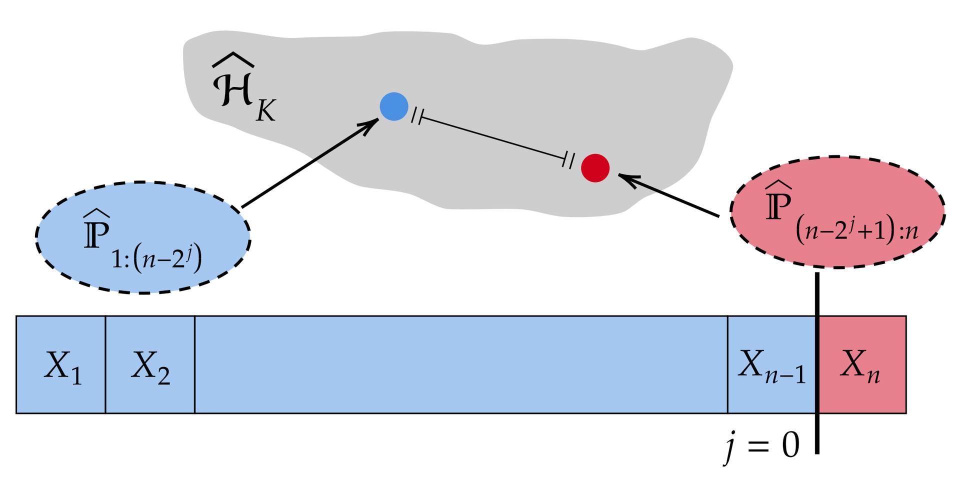
leading to

$$\mathrm{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] = \left\| \mu_{\hat{K}}\left(\hat{\mathbb{P}}_n\right) - \mu_{\hat{K}}\left(\hat{\mathbb{Q}}_n\right) \right\|_{\mathcal{H}_{\hat{K}}} = \left\| \frac{1}{n} \sum_{i=1}^n \hat{z}_K(X_i) - \frac{1}{m} \sum_{i=1}^m \hat{z}_K(Y_i) \right\|_2.$$

- Observation: Both sums can be stored explicitly and MMD can be efficiently computed.

Proposed Change Detection Algorithm: Online RFF-MMD

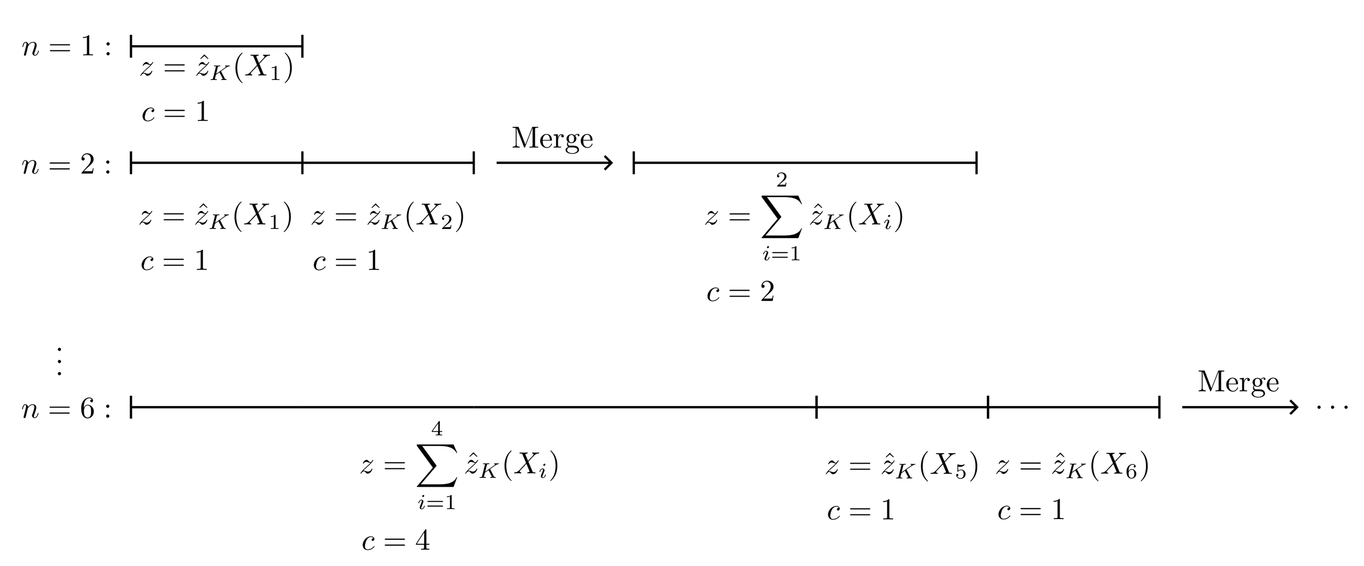
- At the n-th iteration, the algorithm considers $\log_2(n)$ splits of the data stream $\{X_1, \dots, X_n\}$. For every such split the RFF-MMD between empirical measures of the two samples is computed. The process is stopped at the first n for which at least one statistic is larger than a given threshold.



- Formally, the Online RFF-MMD stopping time is defined as

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

- The algorithm has logarithmic time complexity per observation and overall logarithmic space complexity, as illustrated by the following example: upon observing the first $n = 6$ elements $X_1, \dots, X_6$, the algorithm operates as follows



Theoretical Guarantees

With an appropriately chosen sequence of thresholds, RFF-MMD can be made to attain, respectively, a desired average run length or a desired uniform false alarm probability.

Average Run Length

For any $\gamma > 1$, if the sequence of thresholds satisfies

$$\lambda_n \geq \sqrt{2} + \sqrt{2 \log\left(4\gamma \log_2(2\gamma)\right)}$$

for all $n \in \mathbb{N}$, it holds that $\mathbb{E}_\infty[N] \geq \gamma$.

False Alarm Probability

For any $\alpha \in (0, 1)$, if the sequence of thresholds satisfies

$$\lambda_n \geq \sqrt{2} + \sqrt{2\left(\log(n/\alpha) + 2\log\log_2(n) + \log\log_2(2n)\right)}$$

for each $n \in \mathbb{N}$, it holds that $\mathbb{P}_\infty(N < \infty) \leq \alpha$.

With high probability, provided the number of RFFs is chosen sufficiently large, the detection delay is bounded from above by a quantity depending only on the chosen α, the number of pre-change observations, and the squared MMD between the pre- and post-change distributions.

Detection Delay

If λ_n is chosen to control the false alarm probability at some level $\alpha \in (0, 1)$, $\mathrm{supp}(\mathbb{P}) \cup \mathrm{supp}(\mathbb{Q}) \subseteq \mathcal{X}$ for some compact set $\mathcal{X} \subset \mathbb{R}^d$, the quantities η, α, and $\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]$ jointly satisfy

$$\eta \geq C_1 \frac{\log(2\eta/\alpha)}{\left(\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]\right)^2},$$

and the number of random features is chosen so that

$$\sqrt{r} \geq C_2 \frac{C_3 + \sqrt{2\log(2/\alpha)}}{\left(\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]\right)^2},$$

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

$$(N - \eta)^+ \leq 1 \vee C_4 \frac{\log(2\eta/\alpha)}{\left(\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]\right)^2},$$

where C_1, C_2, C_3, and C_4 are absolute constants independent of η, α, and $\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]$.

The detection delay of RFF-MMD is optimal from a minimax perspective, up to logarithmic terms.

Information Theoretic Bounds

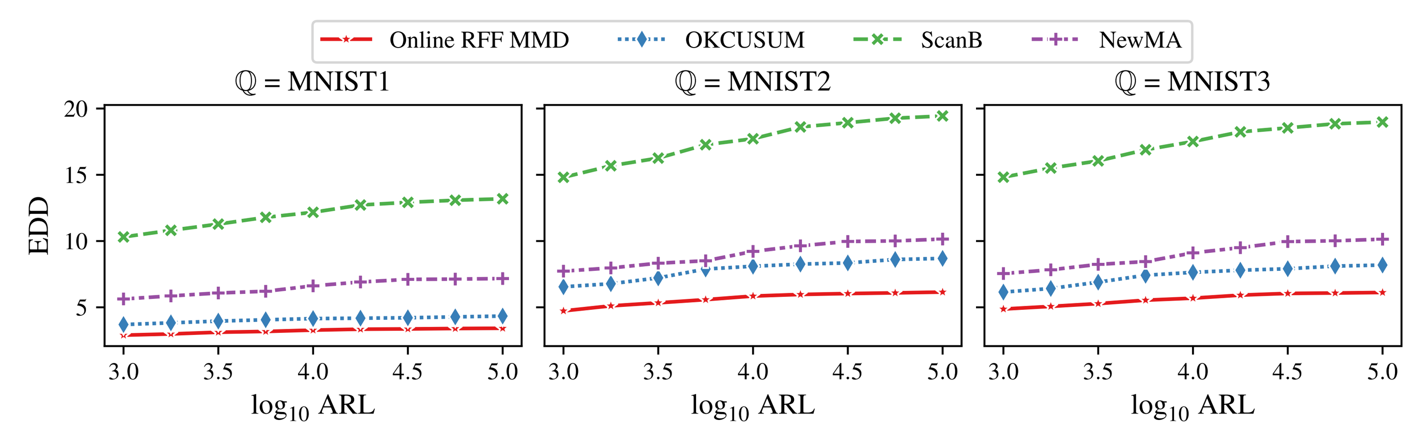
For every bounded, continuous, and translation invariant kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ there is a constant C_K depending only on K and absolute constants $\alpha_0, \beta_0 \in (0, 1)$ independent of K, such that for any $\alpha \leq \alpha_0$ it holds that

$$\inf_{N : \mathbb{P}_\infty(N < \infty) \leq \alpha} \sup_{\substack{\eta > 1 \\ \mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+}} \mathbb{P}_\eta\left(N \geq \eta + C_K \frac{\log(1/\alpha)}{\left(\mathrm{MMD}_K[\mathbb{P}, \mathbb{Q}]\right)^2}\right) \geq \beta_0$$

with the infimum being over all extended stopping times.

Numerical Experiments

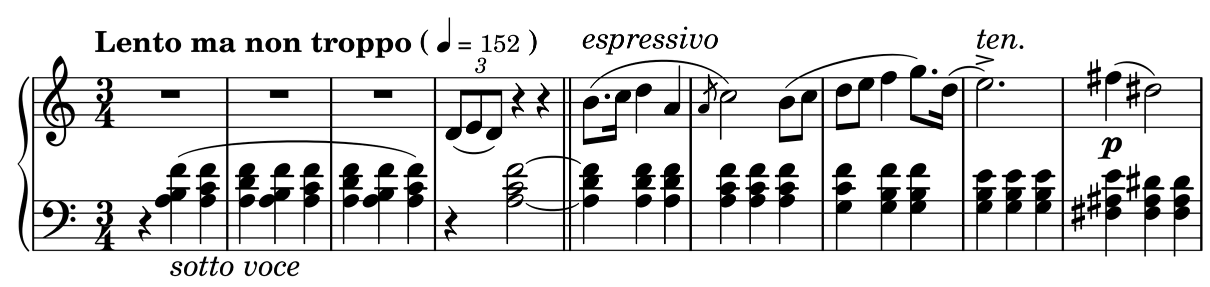
- **MNIST ($d = 768$).** Pre-change: 64 samples of digit 0; post-change: samples of indicated digit.



- **HASC (Human Activity Sensing Consortium; $d = 3$).** Pre-change: 100 samples of "Walking"; post-change: 100 samples of "Staying".

Algorithm	Average delay	Too early	Miss
Online RFF MMD	21.86	2	1
NewMA	34.25	1	5
ScanB	31.20	0	0
OKCUSUM	17.44	1	0
RuLSIF	20.38	2	0

- **Loudness of Chopin's Mazurka Op. 17 No. 4 ($d = 1$).** Change points in loudness information of Chopin's Mazurkas correspond to score positions having dynamic markings, tempo, or expression markings, among others [2]. Our proposed method flags 10 change points too early, and, on the remaining 15 has an average detection delay of 73.67, with a median detection delay of 64.0.



References

[1] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

[2] K. Kosta, R. Killick, O. Bandtlow, and E. Chew. Dynamic change points in music audio capture dynamic markings in score. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[3] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.

[4] B. K. Sriperumbudur and Z. Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1152, 2015.