
Optimal Online Change Detection via Random Fourier Features

Florian Kalinke*

Information Systems
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
florian.kalinke@kit.edu

Shakeel Gavioli-Akilagun*

Department of Statistics
London School of Economics (LSE)
London, UK
s.a.gavioli-akilagun@lse.ac.uk

Abstract

This article studies the problem of online non-parametric change point detection in multivariate data streams. We approach the problem through the lens of kernel-based two-sample testing and introduce a sequential testing procedure based on random Fourier features, running with logarithmic time complexity per observation and with overall logarithmic space complexity. The algorithm has two advantages compared to the state of the art. First, our approach is genuinely online, and no access to training data known to be from the pre-change distribution is necessary. Second, the algorithm does not require the user to specify a window parameter over which local tests are to be calculated. We prove strong theoretical guarantees on the algorithm’s performance, including information-theoretic bounds demonstrating that the detection delay is optimal in the minimax sense. Numerical studies on real and synthetic data show that our algorithm is competitive with respect to the state of the art.

1 Introduction

In the online change point detection problem, data is observed sequentially, and the goal is to flag a change if the distribution of the data changes. The problem dates back to the early work of Page [39], and has now been extensively studied in the statistics and machine learning literature [27, 3, 48]. However, these classical approaches assume that the data are low-dimensional and that the pre- and post-change distributions belong to a known parametric family. In modern applications, both assumptions are usually not satisfied. Examples of modern online change point detection problems include: detecting changes in audio streams [4], in videos [1, 24], in highway traffic data [14], in internet traffic data [28], or in cardiac time series [56]. Further, such data frequently has high volume, and online change point detection procedures should still be able to process new data in real time and with limited memory.

While algorithms for the problem of non-parametric online change point detection have been proposed—see Section 2 for a brief overview—state-of-the-art methods suitable for modern data suffer from at least one of the following two limitations. First, most procedures are not genuinely online from a statistical perspective, in the sense that they either assume the pre-change distribution to be known completely or they assume having access to historical data known to be from the pre-change distribution. Second, most approaches require the user to specify a window parameter over which local tests for a change in distribution will be applied. Choosing such a window is notoriously difficult [43], and choosing the window too small or too large leads to a reduction in power or an increase in the detection delay, respectively.

*Contributed equally.

Motivated by the above limitations and challenges, we propose a new algorithm called Online RFF-MMD (Random Fourier Feature Maximum Mean Discrepancy). On a high level, the algorithm performs sequential two-sample tests based on the kernel-based maximum mean discrepancy (MMD; [49, 13]). Crucially, approximating the maximum mean discrepancy using random Fourier features (RFFs; [41]) leads to a detection statistic that can be computed in linear and updated in constant time. By embedding these local tests in a sequential testing scheme on a dyadic grid of candidate change point locations, we obtain an algorithm that does not require a window parameter and features a time and space complexity logarithmic in the amount of data observed.

This article makes the following contributions:

- **Computational efficiency:** We propose Online RFF-MMD, a fully non-parametric change point detection algorithm that does not require access to historical data, has no window parameter, and features logarithmic runtime and space complexity.
- **Minimax optimality:** Online RFF-MMD comes with strong theoretical guarantees. In particular, we derive information-theoretic bounds showing that the detection delay incurred by Online RFF-MMD is optimal up to logarithmic terms in the minimax sense. While related results are known in the offline setting [37, 38] and in the parametric online setting [27, 59], ours is the first result of this kind for kernel-based online change point detection.
- **Empirical validation:** We perform a suite of benchmarks on synthetic data and on the MNIST data set to demonstrate the applicability of the proposed method. Our approach achieves competitive results throughout all experiments.

The article is structured as follows. We recall related work in Section 2 and introduce our notations and the problem in Section 3. Section 4 presents our algorithm and its guarantees, and Section 5 its minimax optimality. Experiments are in Section 6 and limitations in Section 7. Additional results and all proofs are in the supplement.

Table 1: Comparison of kernel-based change detectors. Genuinely online — whether the algorithm can be executed without reference data known to be from the pre-change distribution; Window free — whether the algorithm requires selection of a window parameter over which the detection statistic is calculated; Time comp. — runtime complexity per new observation; Space comp. — total space complexity; n — total number of observations.

Algorithm	Genuinely online	Window free	Time comp.	Space comp.
Scan B -statistics [30]	✗	✗	$\mathcal{O}(NW^2)$	$\mathcal{O}(NW)$
NEWMA [23]	✓	✗	$\mathcal{O}(r)$	$\mathcal{O}(r)$
Online kernel CUSUM [55]	✗	✗	$\mathcal{O}(NW^2)$	$\mathcal{O}(NW)$
Online RFF-MMD	✓	✓	$\mathcal{O}(r \log n)$	$\mathcal{O}(r \log n)$

2 Related work

In the case of univariate data, numerous methods for non-parametric online change point detection have been proposed; we refer to Ross [45] for a more comprehensive overview. The most successful among these approaches exploit the fact that all information is contained in the data’s empirical distribution function, which is a functional of the ranks. Rank-based online change point detection methods have been proposed by [12, 18, 44]. However, their extension to multivariate data is challenging as this requires a computationally efficient multivariate analogue to scalar ranks.

To tackle change point detection on multivariate data, many approaches exist; see Wang and Xie [54] for a survey. For example, Yilmaz [57], Kurt et al. [25] introduce procedures using summary statistics based on geometric entropy minimization, but assume knowledge of the pre-change distribution. An alternative non-parametric approach is using information on distances between data points [6, 7]. However, here one can construct alternatives which the procedures will always fail to detect as the limits of the test statistics employed do not metricize the space of probability distributions.

One principled approach to tackle this challenging setting is using kernel-based two sample tests via the MMD, which we recall briefly in Section 3.2. The key property, if the underlying kernel is characteristic [10, 51], is that the MMD metricizes the space of probability distributions and thus

allows detecting any change. However, the sample estimators of the MMD typically have a quadratic runtime complexity, which prohibits their direct application in the online setting. To overcome this challenge, Zaremba et al. [61] propose Scan B -statistics, which achieve sub-quadratic time complexity by splitting the data into blocks and computing the quadratic-time MMD on each block. Consequently, Li et al. [29, 30] introduce an online change point detection algorithm that recursively estimates the Scan B -statistic over a sliding window. Extending this idea, Wei and Xie [55] propose an algorithm which performs the same operation over grid of windows of different sizes. However, these algorithms all require the choice of a window parameter and are not genuinely online as they require historical data known to be from the pre-change distribution. Keriven et al. [23] propose comparing two exponentially smoothed MMD statistics, where the MMD is approximated by using random Fourier features. However, the window selection problem is not avoided because, as shown by the authors, the smoothed statistic can be interpreted as computing differences between MMDs calculated on two windows of different sizes.

We summarize the kernel-based approaches coming with theoretical guarantees in Table 1 and note that we present an extension of our proposed Online RFF-MMD approach that allows taking historical data into account in the supplement.²

3 Preliminaries

In this section, we formally introduce the online change point detection problem (Section 3.1), and recall kernel-based two sample testing together with its RFF-based approximation (Section 3.2).

3.1 Problem statement

We consider a data stream X_1, X_2, \dots observed online, where the X_t -s are independent random variables taking values in \mathbb{R}^d , with d arbitrary but fixed. Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathbb{R}^d) := \mathcal{M}_1^+(\mathbb{R}^d)$ and $\mathbb{P} \neq \mathbb{Q}$, where $\mathcal{M}_1^+(\mathbb{R}^d)$ denotes the set of all Borel probability measures on \mathbb{R}^d . We assume that there exists an $\eta \in \mathbb{N} := \{1, 2, \dots\}$ such that

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

The goal is to stop the process with minimal delay as soon as η is reached, but not before. Note that we may have $\eta = \infty$ in which case the process should never be stopped. Formally, one wants to test

$$H_{0,n} : X_t \sim \mathbb{P} \text{ for each } t \leq n \text{ and some } \mathbb{P} \in \mathcal{M}_1^+ \text{ versus} \\ H_{1,n} : \exists \eta < n \text{ s.t. } X_t \sim \begin{cases} \mathbb{P} & \text{if } 1 \leq t \leq \eta \\ \mathbb{Q} & \text{if } \eta < t \leq n \end{cases}, \text{ and some } \mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+ \text{ where } \mathbb{P} \neq \mathbb{Q}, \quad (1)$$

for each $n \in \mathbb{N}$, until a local null is rejected. A secondary aim, once a local null has been rejected, is to accurately estimate η . Solving the aforementioned problems boils down to constructing an extended stopping time N , which, in a sense we make precise in Section 4.3, is close to η . Let $\mathcal{F}_t = \sigma(X_s \mid s = 1, \dots, t)$ be the natural filtration generated by the X_s 's up to time t and recall that a random variable N is an extended stopping time if (i) N takes values in $\mathbb{N} \cup \{\infty\}$ and (ii) for each $t \in \mathbb{N}$ the event $\{N \leq t\}$ is \mathcal{F}_t -measurable.

Minimizing the distance between η and N is analogous to maximizing the power of a particular sequential testing procedure, and it is therefore natural to impose some conditions on the sequential testing analogue of statistical size; we recall the two most frequent ones in the following. In the sequel, let \mathbb{P}_k be the joint distribution of $\{X_t\}_{t>0}$ when $\{\eta = k\}$ ($k \in \mathbb{N}$), and let \mathbb{E}_k be the expectation under this distribution.

1. The average run length until a spurious rejection under the global null should be bounded from below by a chosen quantity [32]. Specifically, for a given $\gamma > 1$ it should hold that

$$\mathbb{E}_\infty[N] \geq \gamma. \quad (2)$$

²In Table 1, where applicable, r denotes the number of random Fourier features, W denotes the size of the window, and N denotes the number of blocks of historical data. We treat the dimension of the data as fixed.

2. The uniform false alarm probability should be bounded from above by a chosen quantity [27]. Specifically, for a given $\alpha \in (0, 1)$ it should hold that

$$\mathbb{P}_\infty(N < \infty) \leq \alpha. \quad (3)$$

We show in Section 4.3 that Online RFF-MMD is able to satisfy either of the conditions (2) or (3).

3.2 Fast kernel-based two sample tests

To resolve the change point detection problem (3.1), we embed fast two sample tests based on RFF approximations to the MMD into a particular sequential testing scheme. In this section, we briefly recall the MMD statistic and its RFF approximation.

Let \mathcal{H}_K be a reproducing kernel Hilbert space (RKHS; [2, 52]) on \mathbb{R}^d with (reproducing) kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Denote by $\text{supp}(\Lambda) = \overline{\{A \in \sigma(\mathbb{R}^d) \mid \Lambda(A) > 0\}}$ the support of a measure Λ on \mathbb{R}^d , where \overline{A} denotes the closure of a set A . Throughout the article, we make the following assumption on the kernel:

Assumption 1. *The kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is non-negative, continuous, bounded ($\exists B > 0$ s.t. $\sup_{\mathbf{x} \in \mathbb{R}^d} K(\mathbf{x}, \mathbf{x}) \leq B$), translation-invariant ($K(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$ for some positive definite $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$), and characteristic ($\text{supp}(\Lambda) = \mathbb{R}^d$ with $\psi(\mathbf{x}) = \int e^{-i\omega^\top \mathbf{x}} d\Lambda(\omega)$).*

To simplify exposition, we also assume that $K(0, 0) = 1$, which can be achieved by scaling any bounded kernel. The conditions in Assumption 1 are satisfied by a number of commonly used kernels, including the Gaussian, Laplace, and B-spline kernels [51, Table 2].

To any $\mathbb{P} \in \mathcal{M}_1^+$, one can associate the kernel mean embedding $\mu_K(\mathbb{P}) \in \mathcal{H}_K$, taking the form³

$$\mu_K(\mathbb{P}) = \int_{\mathbb{R}^d} K(\cdot, \mathbf{x}) d\mathbb{P}(\mathbf{x}), \quad (4)$$

where the integral is meant in Bochner's sense [9, Chapter II.2]. The continuity and boundedness assumptions ensure the existence of $\mu_K(\mathbb{P})$ for any $\mathbb{P} \in \mathcal{M}_1^+$ [51, Proposition 2]. Expression (4) gives rise to the MMD, which quantifies the distance between two measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+$ via the distance between their mean embeddings in terms of the RKHS norm, and takes the form

$$\text{MMD}_K[\mathbb{P}, \mathbb{Q}] = \|\mu_K(\mathbb{P}) - \mu_K(\mathbb{Q})\|_{\mathcal{H}_K}. \quad (5)$$

Crucially, the characteristic assumption ensures that (4) is injective and that the MMD metricizes the space \mathcal{M}_1^+ [51, Theorem 9], implying $\text{MMD}_K[\mathbb{P}, \mathbb{Q}] = 0$ iff. $\mathbb{P} = \mathbb{Q}$. Given samples $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ and $Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}$ with associated empirical measures $\hat{\mathbb{P}}_n =: X_{1:n}$ and $\hat{\mathbb{Q}}_m =: Y_{1:m}$, respectively, the squared plug-in estimator of $\text{MMD}_K[\mathbb{P}, \mathbb{Q}]$ takes the form

$$\begin{aligned} (\text{MMD}_K[X_{1:n}, Y_{1:m}])^2 &= \left\| \mu_K(\hat{\mathbb{P}}_n) - \mu_K(\hat{\mathbb{Q}}_m) \right\|_{\mathcal{H}_K}^2 = \left\| \frac{1}{n} \sum_{i=1}^n K(\cdot, X_i) - \frac{1}{m} \sum_{i=1}^m K(\cdot, Y_i) \right\|_{\mathcal{H}_K}^2 \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K(X_i, X_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m K(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m K(X_i, Y_j). \end{aligned} \quad (6)$$

The computation of (6) costs $\mathcal{O}((\max(m, n))^2)$, rendering its use in an online testing procedure computationally infeasible.

Random Fourier features [41, 50] alleviate this bottleneck in the offline setting; we recall the method in the following. For some $\omega \in \mathbb{R}^d$ write $\zeta_\omega(\mathbf{x}) = e^{i\omega^\top \mathbf{x}}$, where $i = \sqrt{-1}$. By Bochner's theorem,

$$K(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{R}^d} e^{i\omega^\top (\mathbf{x} - \mathbf{y})} d\Lambda(\omega) = \mathbb{E}_{\omega \sim \Lambda} [\zeta_\omega(\mathbf{x}) \zeta_\omega(\mathbf{y})^*], \quad (7)$$

where $*$ denotes the complex conjugate and, using that $K(0, 0) = 1$, one has that $\Lambda \in \mathcal{M}_1^+$. As Λ and K are real-valued, Euler's identity implies that $\int e^{i\omega^\top (\mathbf{x} - \mathbf{y})} d\Lambda(\omega) = \int \cos(\omega^\top (\mathbf{x} - \mathbf{y})) d\Lambda(\omega)$.

³For $\mathbf{x} \in \mathbb{R}^d$, $K(\cdot, \mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ denotes the map $\mathbf{x}' \mapsto K(\mathbf{x}', \mathbf{x})$.

Therefore, using that $\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta$, picking some $r \in \mathbb{N}$, and sampling $\omega_1, \dots, \omega_r \stackrel{\text{i.i.d.}}{\sim} \Lambda$, a low variance estimator for $K(\mathbf{x}, \mathbf{y})$ is given by

$$\hat{K}(\mathbf{x}, \mathbf{y}) := \langle \hat{z}_K(\mathbf{x}), \hat{z}_K(\mathbf{y}) \rangle, \text{ where } \hat{z}_K(\mathbf{x}) = \frac{1}{\sqrt{r}} \left((\sin(\omega_j^\top \mathbf{x}), \cos(\omega_j^\top \mathbf{x})) \right)_{j=1}^r \in \mathbb{R}^{2r}. \quad (8)$$

By noting that $\hat{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel associated with the RKHS $\mathcal{H}_{\hat{K}} = \mathbb{R}^{2r}$, we may approximate (6) by

$$\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] = \left\| \mu_{\hat{K}}(\hat{\mathbb{P}}_n) - \mu_{\hat{K}}(\hat{\mathbb{Q}}_m) \right\|_{\mathcal{H}_{\hat{K}}} = \left\| \frac{1}{n} \sum_{i=1}^n \hat{z}_K(X_i) - \frac{1}{m} \sum_{i=1}^m \hat{z}_K(Y_i) \right\|_2. \quad (9)$$

Importantly, as the mean embeddings in (9) are Euclidean vectors, their distance can be computed with the standard Euclidean norm. This leads to a statistic which can be computed in linear time and updated in constant time, allowing its use for sequential testing.

4 Online change point detection via random Fourier features

In this section, we present our proposed stopping time for resolving the change point detection problem. In particular, we give a precise definition in Section 4.1, an efficient algorithm in Section 4.2, and theoretical guarantees in Section 4.3.

4.1 The RFF-MMD stopping time

The intuitive construction of our RFF-MMD stopping time is as follows. We begin by choosing an $r \in \mathbb{N}$ and a kernel K , and construct its RFF approximation (8) using r random features. For every $n \geq 2$, having observed data $\{X_1, \dots, X_n\}$, we consider $\log_2 n$ possible sample splits of the domain $\{1, \dots, n\}$ at locations $n - 2^j$ with $j = 0, \dots, \lfloor \log_2 n \rfloor - 1$. For every such split, we approximate the MMD between the two samples using (9). A change is declared at the first n for which at least one such statistic, appropriately normalized so that it is $\mathcal{O}_P(1)$ under its local null, is larger than a given threshold. Formally, we have

$$N = \inf \left\{ n \geq 2 \mid \bigcup_{j=1}^{\lfloor \log_2(n) \rfloor - 1} \sqrt{\frac{2^j(n-2^j)}{n}} \text{MMD}_{\hat{K}}[X_{1:(n-2^j)}, X_{(n-2^j+1):n}] > \lambda_n \right\}, \quad (10)$$

where $\{\lambda_n \mid n \in \mathbb{N}\}$ is a non-decreasing sequence that we make precise in Section 4.3, taking requirements (2) or (3) into account.

The first use of an exponential grid in online change point detection appears to be due to Lai [26]. Recently, similar techniques have been used by Yu [58], Kalinke et al. [22], Moen [35]. The dyadic grid used in (10) has two advantages. First, only a logarithmic number of tests must be performed with each new observation. Second, the grid is sufficiently dense, so the obtained stopping time has essentially the same behavior as the computationally infeasible variant, which considers every possible candidate change point location.

4.2 The RFF-MMD algorithm

We now present an efficient implementation of (10) and analyze its runtime and space complexity. We show the pseudo code of our proposed method in Algorithm 1; see also the corresponding figure in the supplement for a visual summary. The details are as follows. For each new observation X_t , we create a new window W , storing $z = \hat{z}_K(X_t)$ and $c = 1$ (Lines 3–4). The window W is then added to the list of all windows \mathcal{W} (Line 5). The remaining algorithm has two main parts.

1. **Change point detection.** To detect changes, we iterate all $|\mathcal{W}| - 1$ dyadic points i (Line 6), and, for each i , merge the feature maps coming before i and coming after i (along with their counts) to compute the MMD statistic (Lines 7–9). If the statistic exceeds the threshold (Line 10; see Section 4.3 for its value), a change is flagged and we drop the data coming before the change.

Algorithm 1 Online RFF-MMD change point detection

Input: Stream X_1, X_2, \dots and a sequence of thresholds $\{\lambda_t \mid t \in \mathbb{N}\}$.

Output: Change point location and detection time.

```
1:  $\mathcal{W} \leftarrow$  empty list
2: for  $X_t \in X_1, X_2, \dots$  do ▷ Main loop
3:    $W.z \leftarrow \hat{z}_K(X_t)$ 
4:    $W.c \leftarrow 1$ 
5:    $\mathcal{W} \leftarrow \mathcal{W}.append(W)$ 
6:   for  $i \in 1, \dots, |\mathcal{W}| - 1$  do ▷ Detect changes
7:      $c_1 \leftarrow \sum_{j=i+1}^{|\mathcal{W}|} \mathcal{W}_j.c$ 
8:      $c_2 \leftarrow \sum_{j=1}^i \mathcal{W}_j.c$ 
9:      $MMD_{\hat{K}} \leftarrow \left\| \frac{1}{c_1} \sum_{j=i+1}^{|\mathcal{W}|} \mathcal{W}_j.z - \frac{1}{c_2} \sum_{j=1}^i \mathcal{W}_j.z \right\|_2$ 
10:    if  $\sqrt{\frac{c_1 c_2}{c_1 + c_2}} MMD_{\hat{K}} \geq \lambda_t$  then
11:      print Change detected at element  $X_t$ ; most likely at position  $i$ .
12:      Drop tail of  $\mathcal{W}$ 
13:    while  $|\mathcal{W}| \geq 2$  do ▷ Maintain exponential structure
14:       $W_1 \leftarrow \text{pop } \mathcal{W}$ 
15:       $W_2 \leftarrow \text{pop } \mathcal{W}$ 
16:      if  $W_1.c = W_2.c$  then
17:         $W.c \leftarrow W_1.c + W_2.c$ 
18:         $W.z \leftarrow W_1.z + W_2.z$ 
19:         $\mathcal{W} \leftarrow \mathcal{W}.append(W)$ 
20:      else
21:         $\mathcal{W} \leftarrow \mathcal{W}.append(W_1).append(W_2)$ 
22:      break
```

2. **Structure maintenance.** To set up and maintain the dyadic structure, we merge windows that have the same counts (Line 16) by first summing their z -s and their c -s (Lines 17–18), and then replacing them in the list of windows \mathcal{W} accordingly (Line 19). We note that pop removed the windows beforehand.

We now analyze the runtime and space complexity of Online RFF-MMD. For each insert operation, Algorithm 1 performs three steps, which we analyze independently.

1. **Setup.** The computation of $\hat{z}_K(X_t)$, defined in (8), requires computing $2r$ trigonometric functions of d -dimensional inner products and thus is in $\mathcal{O}(rd)$.
2. **Change point detection.** The dominating term is computing $MMD_{\hat{K}}$, which requires $\mathcal{O}(|\mathcal{W}|r)$ computations. Repeating the computation $|\mathcal{W}|$ times leads to a cost of $\mathcal{O}(|\mathcal{W}|^2 r)$. The calculation of the threshold is in $\mathcal{O}(1)$, which gives an overall cost of $\mathcal{O}(|\mathcal{W}|^2 r)$. However, we note that memoization of all sums allows to implement the change point 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) and thereby reduces the runtime complexity to $\mathcal{O}(|\mathcal{W}|r)$.
3. **Maintenance.** In the worst case, $\mathcal{O}(|\mathcal{W}|)$ merge operations need to be performed. Each merge requires $\mathcal{O}(r)$ operations, which yields a total cost of $\mathcal{O}(|\mathcal{W}|r)$.

Adding the results obtained in steps 1.–3. shows that the algorithm has an overall runtime complexity of $\mathcal{O}(|\mathcal{W}|r) = \mathcal{O}(r \log n)$ per insert operation. As the algorithm stores, for each $W \in \mathcal{W}$, a number ($W.c$) and a vector ($W.z \in \mathbb{R}^{2r}$), the total space complexity when having observed n samples is $\mathcal{O}(|\mathcal{W}|r) = \mathcal{O}(r \log n)$. We note that r is a fixed parameter in practice and thus constant.

4.3 Theoretical results

In this section, we analyze the theoretical behavior of the RFF-MMD algorithm. We first study the behavior of the stopping time defined in (10) under the global null of no change. Theorems 1 and 2 show that with an appropriately chosen sequence of thresholds the stopping time can be made to attain, respectively, a desired average run length (2) or a desired uniform false alarm probability (3).

Theorem 1. Let N be the extended stopping time defined via (10). For any $\gamma > 1$, if the sequence of thresholds satisfies $\lambda_n \geq \sqrt{2} + \sqrt{2 \log(4\gamma \log_2(2\gamma))}$ for all $n \in \mathbb{N}$, it holds that $\mathbb{E}_\infty[N] \geq \gamma$.

Theorem 2. Let N be the extended stopping time defined via (10). For any $\alpha \in (0, 1)$, if the sequence of thresholds satisfies $\lambda_n \geq \sqrt{2} + \sqrt{2(\log(n/\alpha) + 2 \log(\log_2(n)) + \log(\log_2(2n)))}$ for each $n \in \mathbb{N}$, it holds that $\mathbb{P}_\infty(N < \infty) \leq \alpha$.

We emphasize that these guarantees do not depend on the number of random features used in constructing (8) and the bounds on the threshold sequences do not require any knowledge of the pre-change distribution.

Next, we study the detection delay incurred by (10) when the threshold sequence is chosen to control the uniform false alarm probability at some level $\alpha \in (0, 1)$. In the following we assume that the data take values in a compact subset of \mathbb{R}^d , and denote the Lebesgue measure of a set by $|\cdot|$. The following result shows that with high probability, provided the number of RFFs is chosen sufficiently large, the detection delay incurred by (10) 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.

Theorem 3. Let N be the extended stopping time defined via (10) with threshold sequence $\{\lambda_n \mid n \in \mathbb{N}\}$ defined as in Theorem 2 for a chosen $\alpha \in (0, 1)$. If $\text{supp}(\mathbb{P}) \cup \text{supp}(\mathbb{Q}) \subseteq \mathcal{X}$ for some compact set $\mathcal{X} \subset \mathbb{R}^d$, the quantities η , α , and $\text{MMD}_K[\mathbb{P}, \mathbb{Q}]$ jointly satisfy

$$\eta \geq C_1 \frac{\log(2\eta/\alpha)}{(\text{MMD}_K[\mathbb{P}, \mathbb{Q}])^2}, \quad (11)$$

and the number of random features in (9) is chosen so that

$$\sqrt{r} \geq C_2 \frac{h(d, |\mathcal{X}|, \sigma) + \sqrt{2 \log(2/\alpha)}}{(\text{MMD}_K[\mathbb{P}, \mathbb{Q}])^2}, \quad (12)$$

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

$$(N - \eta)^+ \leq 1 \vee C_3 \frac{\log(2\eta/\alpha)}{(\text{MMD}_K[\mathbb{P}, \mathbb{Q}])^2},$$

where C_1 , C_2 , and C_3 are absolute constants independent of η , α , and $\text{MMD}_K[\mathbb{P}, \mathbb{Q}]$, and, with $\sigma^2 = \int \|\omega\|_2^2 d\Lambda(\omega)$, we have put

$$h(d, |\mathcal{X}|, \sigma) = 23\sqrt{2d \log(2|\mathcal{X}| + 1)} + 32\sqrt{2d \log(\sigma + 1)} + 16\sqrt{2d [\log(2|\mathcal{X}| + 1)]^{-1}},$$

which is likewise independent of η , α , and $\text{MMD}_K[\mathbb{P}, \mathbb{Q}]$.

Condition (11) can be interpreted as a signal strength requirement, measuring the strength according to the number of observations from the pre-change distribution and the squared MMD between \mathbb{P} and \mathbb{Q} . The term $\log(2\eta/\alpha)$ reflects the cost of multiple testing when the data are drawn from \mathbb{P} . Such requirements are unavoidable from the minimax perspective in the corresponding offline problem [58, 38, 37], and the discussion in Yu et al. [59, Section 4.1] suggests that the same is true for genuinely online change point problems.

The requirement on the number of RFFs in (12) depends on $\text{MMD}_K[\mathbb{P}, \mathbb{Q}]$, which is unknown in practice. However, if one assumes an asymptotic setting with a fixed distance between \mathbb{P} and \mathbb{Q} , and $\alpha \downarrow 0$, then (12) suggests that the number of RFFs should be chosen as $r = \Theta(\log 1/\alpha)$. To put this result into perspective, we compare it to online change procedures having a window. Here, the optimal window length also depends on the distance between the pre- and post-change distributions [30, 55]. However, choosing the window larger than this quantity can lead to an increase in the detection delay. This is not the case for RFF-MMD: in practice, one may choose the number of RFFs as large as possible subject to computational constraints, and choosing a larger number of RFFs does not negatively impact the detection delay.

5 Minimax optimality of RFF-MMD

Recall that with the conditions of Assumption 1, the underlying kernel is characteristic and MMD_K metricizes the space \mathcal{M}_1^+ . Therefore, $\text{MMD}_K[\mathbb{P}, \mathbb{Q}] > 0$ for any $\mathbb{P} \neq \mathbb{Q}$, and (3) guarantees that our

stopping time (10) obtains a finite detection delay for any fixed alternative. Still, one may ask whether the detection delay is optimal. The following theorem resolves this question and shows that the detection delay of RFF-MMD is essentially optimal from a minimax perspective, up to logarithmic terms.

Theorem 4. *For every continuous, bounded, and shift invariant kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 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*

$$\inf_{N : \mathbb{P}_\infty(N < \infty) \leq \alpha} \sup_{\substack{\eta > 1 \\ \mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+}} \mathbb{P}_\eta \left(N \geq \eta + C_K \frac{\log(1/\alpha)}{(\text{MMD}_K[\mathbb{P}, \mathbb{Q}])^2} \right) \geq \beta_0 \quad (13)$$

with the infimum being over all extended stopping times.

We remark that in the online change point detection literature [36, 42], it is more common to study the expected risk of a stopping time. For example, for fixed \mathbb{P}, \mathbb{Q} , it is common to work with the so-called worst-worst-case average detection delay [33] of a given stopping time N , which is defined via

$$\sup_{\eta > 1} \text{ess sup } \mathbb{E}_\eta \left[(N - \eta)^+ \mid \mathcal{F}_\eta \right].$$

However, in the absence of further restrictions, studying this quantity for the problem at hand does not seem possible. In fact, as long as $\mathbb{P}^{\otimes \eta} := \mathbb{P} \otimes \dots \otimes \mathbb{P}$ and $\mathbb{Q}^{\otimes \eta} := \mathbb{Q} \otimes \dots \otimes \mathbb{Q}$ have a total variation distance smaller than 1, one can couple the given process and a process where all X_t 's are drawn from \mathbb{Q} so that with non-zero probability the two processes have identical sequences. In this case, we either lose control of the null and α cannot be arbitrarily close to zero, or we maintain control over the null but with non-zero probability the detection delay is infinite.

6 Experiments

We collect our experiments on synthetic data in Section 6.1 and on the MNIST data set in Section 6.2. We refer to the supplement for a numerical comparison of the different thresholds for the stopping rule. To interpret the change point detection performance of the proposed method, we compare its average run length (ARL) and expected detection delay (EDD) to the existing kernel-based methods presented in Table 1.⁴ For all experiments, we use the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}$ with γ set by the median heuristic [11] or its RFF approximation, depending on the algorithm. All results were obtained on a PC with Ubuntu 20.04 LTS, 124GB RAM, and 32 cores with 2GHz each.

6.1 Synthetic data

In this section, we evaluate the runtimes of different configurations of the proposed Online RFF-MMD algorithm and compare its change point detection performance on synthetic data to that of other kernel-based approaches.

Runtime. Figure 1 summarizes the runtime results of Algorithm 1 with random Fourier features $r \in \{10, 50, 100, 500, 1.000\}$ and for streams of length up to $n = 250.000$. The experiments verify the $\mathcal{O}(r \log n)$ runtime complexity of the proposed algorithm, derived analytically in Section 4.2. We note that the dependence on d is linear; we consider $d = 1$ only.

ARL vs. expected detection delay. To illustrate the EDD for a given target ARL, we reproduce the experiments of Wei and Xie [55, Figure 4], also taking our method into account. Specifically, we consider the pre-change distribution $\mathbb{P} = \mathcal{N}(\mathbf{0}_{20}, \mathbf{I}_{20})$, and set the parameters of each algorithm as follows. Matching the settings of the reproduced experiment, we choose $B_{\max} = 50$ and $N = 15$ for online kernel CUSUM; for Scan B-statistics and NewMA, we set $B_0 = 50$. The remaining parameters of NewMA then follow from the heuristics detailed by the authors [23]. For Online RFF-MMD, we set $r = 1.000$. We compute the thresholds for a given target ARL by processing $10 \times (\text{target ARL})$ samples with each algorithm, repeating for 100 Monte Carlo (MC) iterations,

⁴ All code replicating the experiments is available at <https://github.com/FlopsKa/rff-change-detection>.

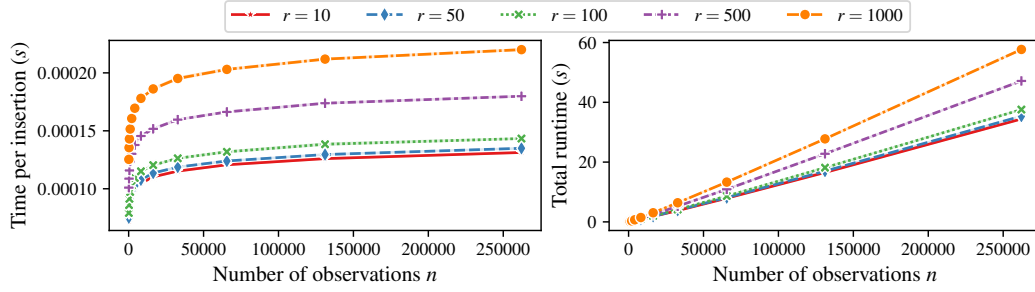


Figure 1: Average runtime (10 repetitions) of RFF-MMD per insert operation (left) and total (right).

and computing the $1 - 1/(\text{target ARL})$ quantile of the resulting test statistics. For approximating the EDD of each algorithm, we draw 64 samples from \mathbb{P} , respectively, before sampling from \mathbb{Q} ; we report the average over 100 repetitions. OKCUSUM and ScanB additionally receive 1.000 samples from \mathbb{P} upfront, to use as a reference sample. For NewMA, we process 400 additional samples from \mathbb{P} for both the MC estimate and the EDD experiment, to reduce its variance.

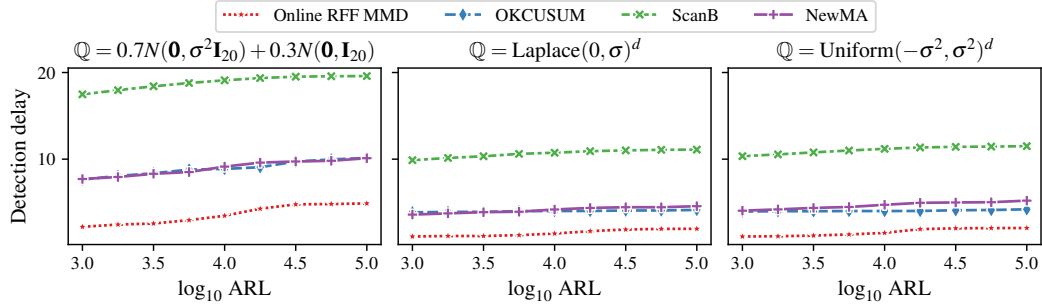


Figure 2: Average detection delay from $\mathbb{P} = \mathcal{N}(\mathbf{0}_{20}, \mathbf{I}_{20})$ to the \mathbb{Q} indicated on top ($d = 20, \sigma = 2$).

Having processed the indicated number of samples from \mathbb{P} , we then start sampling from either a mixed normal, a Laplace, or a Uniform distribution. Figure 2 collects the average detection delay; the respective post-change distribution \mathbb{Q} is given on top. The results show that our algorithm achieves a smaller detection delay than the competitors for all considered post-change distributions, sometimes by a large margin.

6.2 MNIST data

In this section, we interpret the MNIST data set [8] as high-dimensional data stream, similar to Wei and Xie [55, Figure 7], with the goal of detecting a change when the digit changes from 0 to a different digit. The experimental setup matches that of Section 6.1, but with $d = 784$. Figure 3 collects our results; the results for the digits 4–6 are similar and in the supplement. Similar to Figure 2, the proposed Online RFF-MMD algorithm shows very good performance. In particular, it achieves a lower EDD than all tested competitors throughout.

7 Limitations

As with all kernel-based tests, the choice of kernel impacts the power of the test. While our theoretical guarantees (Section 4.3) hold for any kernel satisfying Assumption 1 and do not require any pre-change data, one usually selects the kernel or its parameters using a few available samples in practice. While kernel optimization is not the focus of this work, there exist works [19–21, 31, 46, 47, 15–17] to (approximately) achieve this goal; it is interesting future work to tackle this problem in the sequential setting.

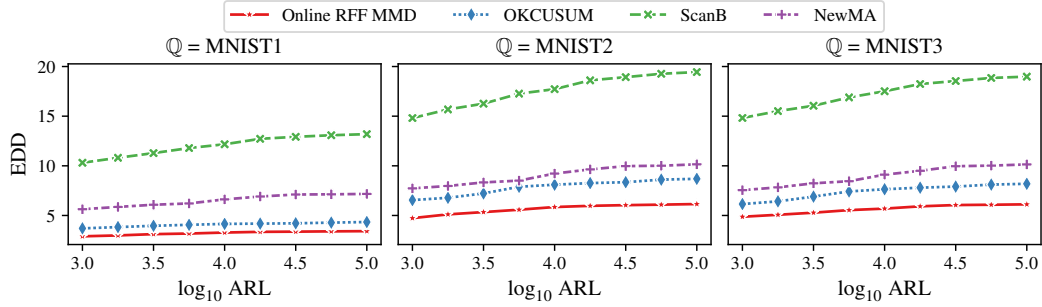


Figure 3: Average detection delay from MNIST digit 0 to digits 1, 2, and 3 (left to right).

Acknowledgments and Disclosure of Funding

The authors thank Zoltán Szabó for helpful discussions. FK thanks Georg Gntuni and Marius Bohnert for exchanges on the algorithm’s implementation. This work was supported by the pilot program Core-Informatics of the Helmholtz Association (HGF).

References

- [1] Abdalbassir Abou-Elailah, Valérie Gouet-Brunet, and Isabelle Bloch. Detection of abrupt changes in spatial relationships in video sequences. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 89–106, 2015.
- [2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.
- [3] Michèle Basseville and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall, 1993.
- [4] Alberto Bietti, Francis R. Bach, and Arshia Cont. An online EM algorithm in hidden (semi-) Markov models for audio segmentation and clustering. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1881–1885, 2015.
- [5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, 2013.
- [6] Hao Chen. Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407, 2019.
- [7] Lynna Chu and Hao Chen. Sequential change-point detection for high-dimensional and non-Euclidean data. *IEEE Transactions on Signal Processing*, 70:4498–4511, 2022.
- [8] Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, pages 141–142, 2012.
- [9] Joseph Diestel and John J. Uhl. *Vector Measures*. American Mathematical Society, 1977.
- [10] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 489–496, 2007.
- [11] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. Technical report, 2018. <https://arxiv.org/abs/1707.07269>.
- [12] Louis Gordon and Moshe Pollak. An efficient sequential nonparametric scheme for detecting a change of distribution. *The Annals of Statistics*, 22(2):763–804, 1994.
- [13] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

- [14] Robert Grossman, Michal Sabala, Anushka Aanand, Steve Eick, Leland Wilkinson, Pei Zhang, John Chaves, Steve Vejck, John Dillenburg, Peter Nelson, et al. Real time change detection and alerts from highway traffic data. In *ACM/IEEE Conference on Supercomputing (SC)*, pages 69–69, 2005.
- [15] Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024.
- [16] Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li. Spectral regularized kernel goodness-of-fit tests. *Journal of Machine Learning Research*, 25(309):1–52, 2024.
- [17] Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 2025. (accepted; preprint: <https://arxiv.org/abs/2404.08278>).
- [18] Douglas M. Hawkins and Qiqi Deng. A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173, 2010.
- [19] Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 181–189, 2016.
- [20] Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning (ICML)*, pages 1742–1751, 2017.
- [21] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 262–271, 2017.
- [22] Florian Kalinke, Marco Heyden, Georg Gntuni, Edouard Fouché, and Klemens Böhm. Maximum mean discrepancy on exponential windows for online change detection. *Transactions on Machine Learning Research*, 2025.
- [23] Nicolas Keriven, Damien Garreau, and Iacopo Poli. NEWMA: A new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020.
- [24] Albert Y. Kim, Caren Marzban, Donald B. Percival, and Werner Stuetzle. Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, 89(12):2529–2536, 2009.
- [25] Mehmet N. Kurt, Yasin Yilmaz, and Xiaodong Wang. Real-time nonparametric anomaly detection in high-dimensional settings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2463–2479, 2020.
- [26] Tze Leung Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(4):613–644, 1995.
- [27] Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.
- [28] Céline Lévy-Leduc and François Roueff. Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics*, 3(2):637–662, 2009.
- [29] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3366–3374, 2015.
- [30] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan B -statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019.
- [31] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning (ICML)*, pages 6316–6326, 2020.

- [32] Gary Lorden. On excess over the boundary. *Annals of Mathematical Statistics*, 41:520–527, 1970.
- [33] Gary Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42:1897–1908, 1971.
- [34] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.
- [35] Per August Jarval Moen. A general methodology for fast online changepoint detection. Technical report, 2025. <https://arxiv.org/abs/2504.09573>.
- [36] George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.
- [37] Carlos Misael Madrid Padilla, Haotian Xu, Daren Wang, Oscar Hernan Madrid Padilla, and Yi Yu. Change point detection and inference in multivariate non-parametric models under mixing conditions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [38] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory*, 68(3):1922–1944, 2021.
- [39] Ewan S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.
- [40] Moshe Pollak and David Siegmund. Sequential detection of a change in a normal mean when the initial value is unknown. *The Annals of Statistics*, 19(1):394–416, 1991.
- [41] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2007.
- [42] Ya’acov Ritov. Decision theoretic optimality of the CUSUM procedure. *The Annals of Statistics*, 18(3):1464–1469, 1990.
- [43] Gaetano Romano, Idris A. Eckley, Paul Fearnhead, and Guillem Rigaill. Fast online changepoint detection via functional pruning CUSUM statistics. *Journal of Machine Learning Research*, 24 (81):1–36, 2023.
- [44] Gaetano Romano, Idris A. Eckley, and Paul Fearnhead. A log-linear non-parametric online changepoint detection algorithm based on functional pruning. *IEEE Transactions on Signal Processing*, 72:594–606, 2024.
- [45] Gordon J. Ross. Parametric and nonparametric sequential change detection in R: The cpm package. *Journal of Statistical Software*, 66:1–20, 2015.
- [46] Antonin Schrab, Benjamin Guedj, and Arthur Gretton. KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 32624–32638, 2022.
- [47] Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18793–18807, 2022.
- [48] David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer, 2013.
- [49] Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.
- [50] Bharath K. Sriperumbudur and Zoltán Szabó. Optimal rates for random Fourier features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1144–1152, 2015.
- [51] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.
- [52] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

- [53] Nicolas Verzelen, Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. Optimal change-point detection and localization. *The Annals of Statistics*, 51(4):1586–1610, 2023.
- [54] Haoyun Wang and Yao Xie. Sequential change-point detection: Computation versus statistical performance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(1):e1628, 2024.
- [55] Song Wei and Yao Xie. Online kernel CUSUM for change-point detection. Technical report, 2022. <https://arxiv.org/abs/2211.15070>.
- [56] Ping Yang, Guy Dumont, and John M. Ansermino. Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE Transactions on Biomedical Engineering*, 53(11):2211–2219, 2006.
- [57] Yasin Yilmaz. Online nonparametric anomaly detection based on geometric entropy minimization. In *IEEE International Symposium on Information Theory (ISIT)*, pages 3010–3014, 2017.
- [58] Yi Yu. A review on minimax rates in change point detection and localisation. Technical report, 2020. <https://arxiv.org/abs/2011.01857>.
- [59] Yi Yu, Oscar Hernan Madrid Padilla, Daren Wang, and Alessandro Rinaldo. A note on online change point detection. *Sequential Analysis*, 42(4):438–471, 2023.
- [60] Vadim Yurinsky. *Sums and Gaussian vectors*. Springer, 1995.
- [61] Wojciech Zaremba, Arthur Gretton, and Matthew B. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Neural Information Processing Systems (NeurIPS)*, pages 755–763, 2013.

A Appendix

This supplementary material is structured as follows. We detail the extension of Online RFF-MMD to take a known or observed pre-change distribution into account in Appendix A.1. In particular, we show that in these settings tighter thresholds are possible. In Appendix A.2, we provide a schematic representation of Algorithm 1. Additional numerical results of MMD-RFF are in Appendix A.3; we include numerical results w.r.t. our tighter thresholds in Appendix A.3.3. Appendix A.4 collects all our proofs.

A.1 Extensions: known or estimable pre-change distribution

In this section, we discuss practical extensions of Online RFF-MMD when additional information about the pre-change distribution is available. Specifically, in Section A.1.1, we show how the algorithm can be adapted to settings in which (i) one has access to historical data known to be from the pre-change distribution, or (ii) one knows the pre-change distribution exactly. In Section A.1.2, we describe how this additional information allows sharpening the thresholds proposed in Theorems 1 and 2 in the main text.

A.1.1 Incorporating information of the pre-change distribution

To begin with, we consider the setting in which, for some $\nu \in \mathbb{N}$, historical data $X_{-\nu+1}, \dots, X_0$ known to be from the (nonetheless unknown) pre-change distribution \mathbb{P} is available. This setting has been studied in the literature from both the parametric [40] and non-parametric [55] perspectives. It is straightforward to extend the Online RFF-MMD stopping time defined in the main text to take advantage of the additional information. Intuitively, for each local test one may prepend the historical data to the block of data take to be from the pre-change distribution. More formally, the following stopping time can be used:

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

This stopping time can be implemented similarly to Algorithm 1, and such an implementation features the same time and space complexity as the original algorithm.

Next, we consider the setting in which the pre-change distribution \mathbb{P} is known exactly. With this additional information, rather than performing local two sample tests, one may perform local one sample tests where the RFF approximation to the mean embedding of the data's empirical distribution is compared to the RFF approximation to the mean embedding of \mathbb{P} . More formally, the following stopping time can be used:

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

The exact knowledge of \mathbb{P} permits a precise approximation or the exact computation of $\mathbb{E}_{X \sim \mathbb{P}}[\hat{z}_K(X) \mid \omega_1, \dots, \omega_r]$, given $\omega_1, \dots, \omega_r$ sampled from Λ . Again, this stopping time can be implemented similarly to Algorithm 1, enjoying the same time and space complexity as the original algorithm.

A.1.2 Sharper thresholds through knowledge of the pre-change distribution

Access to additional information about the pre-change distribution paves the way to sharpening the thresholds proposed in Theorems 1 and 2. Indeed, although Theorem 4 suggests the thresholds proposed in the main paper are unimprovable up to constants, in practice these thresholds will be quite conservative as they are completely agnostic to the distribution of the data.

The main tool for proving Theorems 1 and 2 is Lemma 5, which controls the tail behavior of the test performed by Online RFF-MMD under their respective local nulls. With additional knowledge of the pre-change distribution, Lemma 5 can be significantly sharpened by taking the second moment of the feature map into account. To that end, we have the following result.

Lemma 1. Given two independent samples $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ each with mutually independent entries drawn from some $\mathbb{P} \in \mathcal{M}_1^+$, for any $\varepsilon > 0$, it holds that

$$\mathbb{P}(\text{MMD}_{\tilde{K}}[X_{1:n}, Y_{1:m}] > \varepsilon) \leq 2 \exp \left(-\frac{1}{2} \min(m, n) \varepsilon^2 \left[\tilde{\sigma}^2 + 2\sqrt{2\varepsilon} \right]^{-1} \right),$$

where $\tilde{\sigma}^2 = 2\mathbb{E}_{X \sim \mathbb{P}}[K(X, X)] - \mathbb{E}_{X, Y \sim \mathbb{P}}[K(X, Y)]$

Lemma 1 allows the following improvements of Theorems 1 and 2.

Corollary 1. For any $\gamma > 1$, replacing the thresholds in Theorem 1 with the scale dependent thresholds

$$\begin{aligned} \lambda_{n,j} &= \frac{4\sqrt{2}f(\gamma)}{\sqrt{\min(2^j, n - 2^j)}} + \tilde{\sigma}\sqrt{2f(\gamma)}, \quad n \in \mathbb{N}, j \leq \log_2(n) \\ f(\gamma) &= \log(4\gamma \log_2(2\gamma)) \end{aligned} \quad (14)$$

it holds that $\mathbb{E}_\infty[N] \leq \gamma$ where N is as defined in (10).

Corollary 2. For any $\alpha \in (0, 1)$, replacing the thresholds in Theorem 2 with the scale dependent thresholds

$$\begin{aligned} \lambda_{n,j} &= \frac{4\sqrt{2}f(\alpha, n)}{\sqrt{\min(2^j, n - 2^j)}} + \tilde{\sigma}\sqrt{2f(\alpha, n)}, \quad n \in \mathbb{N}, j \leq \log_2(n) \\ f(\alpha, n) &= \log(n/\alpha) + \log(\log_2(n)) + \frac{1}{2} \log(\log_2(n)) \end{aligned}$$

it holds that $\mathbb{P}_\infty(N < \infty) \leq \gamma$ where N is as defined in (10).

To put the above results in context, we detail Corollary 1. On large scales (large j -s), where detections tend to occur, the scale dependent thresholds behave approximately like $\tilde{\sigma}\sqrt{2\log(4\gamma \log_2(2\gamma))}$, which differs from the threshold used in Theorem 1 by a factor of $\tilde{\sigma}$. Therefore, when $\tilde{\sigma}$ is significantly smaller than 1, the thresholds in (14) will be significantly smaller than those suggested by Theorem 1. As the kernel is assumed to be bounded from above by 1, these smaller thresholds generally occur.

A.2 Further details on the Online RFF-MMD algorithm

Figure 4 illustrates Algorithm 1. The details are as follows. Upon observing the first element X_1 , the algorithm creates a new Window W , storing the feature map $\hat{z}_K(X_1)$ and that it has one element. Similarly, upon observing X_2 , the algorithm creates a new window W' , storing $\hat{z}_K(X_2)$ and $c = 1$. As both windows W, W' have the same counts, the algorithm merges them into a new window W , storing $\hat{z}_K(X_1) + \hat{z}_K(X_2)$ and $c = 2$. The algorithm proceeds in this manner, resulting in the construction of the dyadic grid outlined in Section 4. For example, when observing X_6 , there are again two windows of size 1, which the algorithm merges to store a total of two windows, one capturing X_1, \dots, X_4 and the other one capturing X_5, X_6 . We recall that the observations themselves are never stored explicitly.

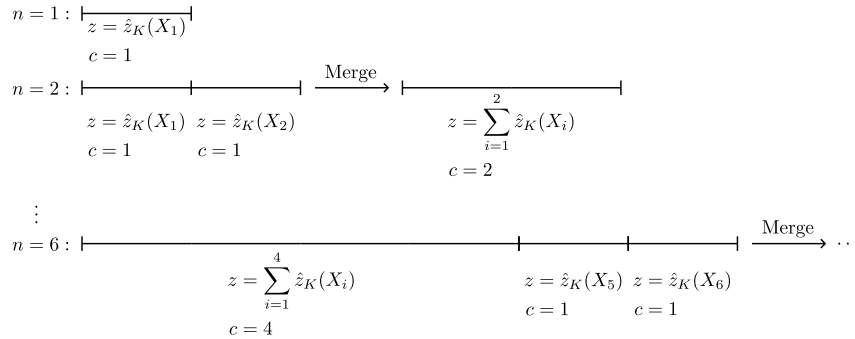


Figure 4: Schematic representation of the proposed algorithm upon observing the first $n = 6$ elements. Merging equal sized “windows” yields the division along dyadic points.

A.3 Additional experiments

In this section, we summarize additional numerical results. The MNIST results with the MC threshold as in the main text are in Section A.3.1. We include results obtained without threshold estimation in Section A.3.2. A comparison of the distribution dependent and distribution-free bounds is in Section A.3.3.

A.3.1 MNIST digits 4–9

In the following Figure 5, we collect the change detection performances of the algorithms of Table 1 on MNIST data, where the digit changes from 0 to one of 4–9; the results on the other digits and the experimental setup are in Section 6.2. As in the corresponding experiment in the main part of this article, our proposed Online RFF-MMD algorithm consistently achieves the lowest expected detection delay, highlighting its good practical performance.

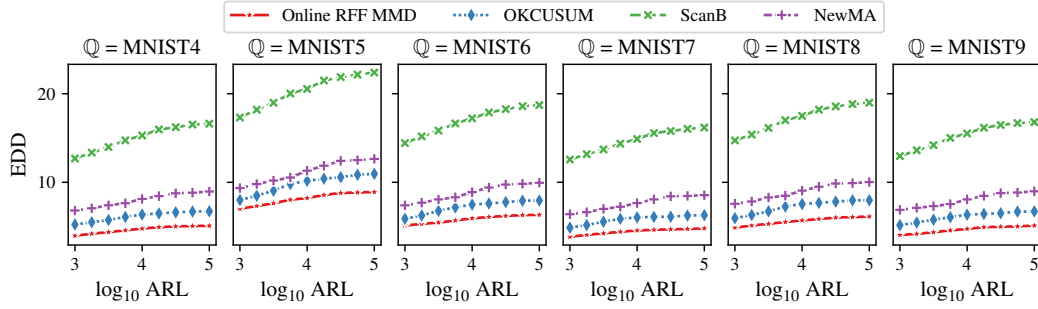


Figure 5: Average detection delay from MNIST digit 0 to digits 4–9 (left to right).

A.3.2 Distribution-free bound

In this section, we show the change detection performance of the proposed Online RFF-MMD algorithm if no pre-change sample is used to estimate the threshold. Instead, we use Theorem 1 to compute the distribution-free threshold sequence $\{\lambda_n \mid n \in \mathbb{N}\}$ for a given target ARL. To obtain an EDD estimate, we sample and process 512 observations from MNIST digit 0 (pre-change) and 1.024 samples from digits 1–9 (post-change), respectively, averaging the detection delay over 100 repetitions. The results are in Figure 6. When comparing to Figure 3 and Figure 5, the figure shows that our method has an increased detection delay, which is due to the looser distribution-free bound (see Section A.3.3 for a numerical comparison). Still, except for the change to the digit 5 with a guaranteed ARL of 10^5 , Online RFF-MMD detects all changes reliably.

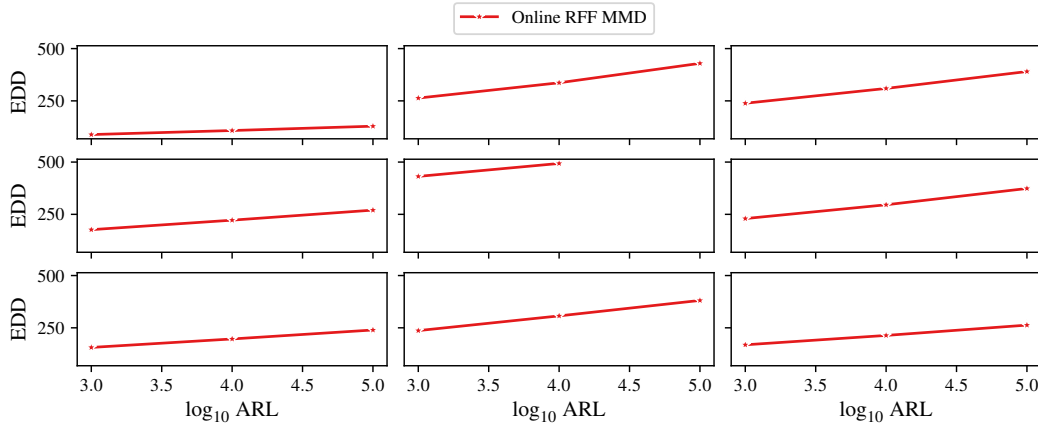


Figure 6: Average detection delay from MNIST digit 0 to digits 1–3, 4–6, 7–9 (top to bottom) with the distribution-free threshold sequence of Theorem 1.

A.3.3 Threshold comparison

In this section, we compare the tightness of our thresholds in the offline two-sample testing setting. Specifically, we fix the level $\alpha = 0.01$ and let $\mathbb{P} = \mathbb{Q} = \mathcal{N}(0, 1)$. We then approximate the $1 - \alpha$ quantile of $\text{MMD}_{\hat{K}}(\hat{\mathbb{P}}_n, \hat{\mathbb{Q}}_n)$ (with $n = 1,000$, \hat{K} approximating the Gaussian kernel with $r = 1,000$ RFFs, and $\gamma > 0$ set by the median heuristic) by (i) obtaining new samples from \mathbb{P}, \mathbb{Q} and (ii) permuting a fixed sample from \mathbb{P}, \mathbb{Q} for 1,000 rounds. Figure 7 shows the respective histograms and the estimated quantiles along with the thresholds obtained by Lemma 5 and Lemma 5, respectively. As one expects, the figure shows that the variance estimate used in Lemma 5 allows to obtain a tighter bound, where we consider the resampling/permutation-based thresholds as ground truth. We emphasize that independent of the threshold used, the resulting test is consistent against any fixed alternative.

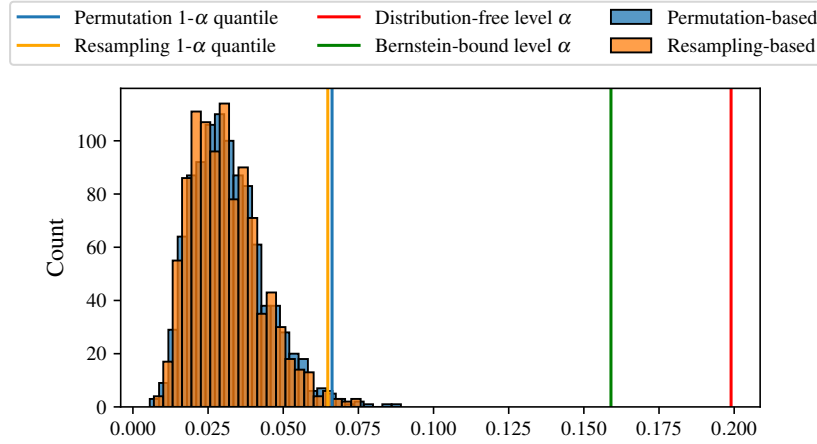


Figure 7: Comparison of different thresholds for the acceptance region of the MMD two-sample test.

A.4 Proofs

This section is dedicated to the proofs of our results stated in the main text. The proof of Theorem 1 is in Appendix A.4.1, that of Theorem 2 is in Appendix A.4.2, that of Theorem 3 is in Appendix A.4.3, and that of our minimax result (Theorem 4) is in Section A.4.4.

A.4.1 Proof of Theorem 1

Proof. For ease of reading, for each $n \geq 2$ and $j = 0, \dots, \lfloor \log_2(n) \rfloor - 1$, put

$$\hat{M}_{n,j} := \sqrt{\frac{2^j(n-2^j)}{n}} \text{MMD}_{\hat{K}}[X_{1:(n-2^j)}, X_{(n-2^j+1):n}]. \quad (15)$$

By the law of total expectation, we have that

$$\begin{aligned} \mathbb{E}_{\infty}[N] &= \mathbb{E}_{\infty}[N \mid N \leq 2\gamma] \mathbb{P}_{\infty}(N \leq 2\gamma) + \mathbb{E}_{\infty}[N \mid N > 2\gamma] \mathbb{P}_{\infty}(N > 2\gamma) \\ &\geq 2\gamma [1 - \mathbb{P}_{\infty}(N \leq 2\gamma)]. \end{aligned} \quad (16)$$

Putting $\lambda = \sqrt{2 \log(4\gamma \log_2(2\gamma))}$, a union bound argument together with Lemma 5 gives that

$$\begin{aligned} \mathbb{P}_{\infty}(N \leq 2\gamma) &= \mathbb{P}_{\infty}\left(\bigcup_{n=2}^{2\gamma} \bigcup_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \hat{M}_{n,j} > \sqrt{2} + \lambda\right) \\ &\leq \sum_{n=1}^{2\gamma} \sum_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \int \dots \int \mathbb{P}_{\infty}(\hat{M}_{n,j} > \sqrt{2} + \lambda_n \mid \omega_1, \dots, \omega_r) d\Lambda(\omega_1) \dots d\Lambda(\omega_r) \\ &\leq 2\gamma \log_2(2\gamma) e^{-\lambda^2/2} = \frac{1}{2}. \end{aligned} \quad (17)$$

Finally, plugging (17) into (16) proves the desired result. \square

A.4.2 Proof of Theorem 2

Proof. Write $\pi_n = \sqrt{2(\log(n/\alpha) + 2\log(\log_2(n)) + \log(\log_2(2n)))}$ for each $n \in \mathbb{N}$. Applying standard peeling arguments [59, 53], we have that

$$\begin{aligned} \mathbb{P}_\infty(N < \infty) &= \mathbb{P}_\infty\left(\bigcup_{n=2}^\infty \bigcup_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \hat{M}_{n,j} > \sqrt{2} + \pi_n\right) \\ &\leq \sum_{l=1}^\infty \mathbb{P}_\infty\left(\bigcup_{n=2^l}^{2^{l+1}} \bigcup_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \hat{M}_{n,j} > \sqrt{2} + \pi_n\right) \\ &\leq \sum_{l=1}^\infty 2^l \max_{2^l \leq n \leq 2^{l+1}} \mathbb{P}_\infty\left(\bigcup_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \hat{M}_{n,j} > \sqrt{2} + \pi_n\right) \end{aligned} \quad (18a)$$

$$\leq \sum_{l=1}^\infty 2^l \max_{2^l \leq n \leq 2^{l+1}} \sum_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \mathbb{P}_\infty\left(\hat{M}_{n,j} > \sqrt{2} + \pi_n\right), \quad (18b)$$

where in line (18a), we apply a union bound and bound the resulting sum by its maximum. Then, applying Lemma 5 as was done in (17), we obtain that

$$\begin{aligned} &\mathbb{P}_\infty(N < \infty) \\ &\leq \sum_{l=1}^\infty 2^l \max_{2^l \leq n \leq 2^{l+1}} \sum_{j=0}^{\lfloor \log_2(n) \rfloor - 1} \int \cdots \int \mathbb{P}_\infty\left(\hat{M}_{n,j} > \sqrt{2} + \pi_n \mid \omega_1, \dots, \omega_r\right) d\Lambda(\omega_1) \cdots d\Lambda(\omega_r) \\ &\leq \sum_{l=1}^\infty l 2^l e^{-(\pi_{2^l})^2/2}. \end{aligned}$$

Finally using the facts that (i) $\exp(-\pi_{2^l}^2/2) \leq \alpha l^{-2}(l+1)^{-1}2^{-l}$ for all $l \in \mathbb{N}$ and (ii) $\sum_{l=1}^\infty l^{-1}(l+1)^{-1} = 1$, we obtain that $\mathbb{P}_\infty(N < \infty) \leq \alpha$. \square

A.4.3 Proof of Theorem 3

Proof. We first observe that for any triplet of integers (m, n, ν) satisfying (i) $m \leq n$ and (ii) $\nu \leq n/2$, given two samples

$$X_{1:n} = \{X_1, \dots, X_{n-\nu}, \tilde{Y}_{n-\nu+1}, \dots, \tilde{Y}_n\} \quad \text{and} \quad Y_{1:m} = \{Y_1, \dots, Y_m\},$$

with mutually independent entries taking values in some bounded set $\mathcal{X} \subset \mathbb{R}^d$ where the X 's are sampled from \mathbb{P} and the Y 's and \tilde{Y} 's are sampled from \mathbb{Q} for some $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ not identical, for any $\varepsilon > 0$, it holds that

$$\begin{aligned} &\mathbb{P}\left(\sqrt{\frac{nm}{n+m}} \text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] \leq \sqrt{2} + \varepsilon\right) \\ &\leq \mathbb{P}\left(2\sqrt{m \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\hat{K}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})|} > \frac{1}{3} \left[\sqrt{m} \left(\frac{\sqrt{2}-1}{2\sqrt{2}} \right) \text{MMD}_K[\mathbb{P}, \mathbb{Q}] - \left(\varepsilon + \frac{10}{\sqrt{2}} \right) \right] \right) \\ &\quad + 4 \exp\left(-\frac{1}{18} \left\{ \left[\sqrt{m} \left(\frac{\sqrt{2}-1}{2\sqrt{2}} \right) \text{MMD}_K[\mathbb{P}, \mathbb{Q}] - \left(\varepsilon + \frac{10}{\sqrt{2}} \right) \right] \vee 0 \right\}^2\right) \end{aligned} \quad (19)$$

where MMD_K is as in (5) and $\text{MMD}_{\hat{K}}$ is as defined in (9). To show (19), let the \tilde{X} 's below be sampled independently from \mathbb{P} and introduce the quantities

$$\begin{aligned} \Delta_1 &= \left\| \frac{1}{n} \left[\sum_{i=1}^{n-\nu} K(\cdot, X_i) + \sum_{i=n-\nu+1}^n K(\cdot, \tilde{X}_i) \right] - \frac{1}{m} \sum_{i=1}^m K(\cdot, Y_i) \right\|_{\mathcal{H}_K} \\ \Delta_2 &= \left\| \frac{1}{\nu} \sum_{i=n-\nu+1}^n K(\cdot, \tilde{X}_i) - \frac{1}{\nu} \sum_{i=n-\nu+1}^n K(\cdot, \tilde{Y}_i) \right\|_{\mathcal{H}_K}. \end{aligned}$$

Note that by the reverse triangle inequality

$$\text{MMD}_K [X_{1:n}, Y_{1:m}] \geq \Delta_1 - \sqrt{\frac{\nu}{n}} \Delta_2.$$

Consequently, using the above and by repeated applications of the triangle inequality, one has that

$$\begin{aligned} & \sqrt{\frac{nm}{n+m}} \text{MMD}_{\hat{K}} [X_{1:n}, Y_{1:m}] \\ & \geq \sqrt{\frac{nm}{n+m}} \left(1 - \sqrt{\frac{\nu}{n}} \right) \text{MMD}_K [\mathbb{P}, \mathbb{Q}] \\ & \quad - \sqrt{\frac{nm}{n+m}} |\text{MMD}_{\hat{K}} [X_{1:n}, Y_{1:m}] - \text{MMD}_K [X_{1:n}, Y_{1:m}]| \end{aligned} \quad (20a)$$

$$- \sqrt{\frac{nm}{n+m}} |\Delta_1 - \mathbb{E}[\Delta_1]| - \sqrt{\frac{2m}{n+m}} \sqrt{\frac{\nu}{2}} |\Delta_2 - \mathbb{E}[\Delta_2]| \quad (20b)$$

$$- \sqrt{\frac{nm}{n+m}} |\mathbb{E}[\Delta_1] - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]| - \sqrt{\frac{2m}{n+m}} \sqrt{\frac{\nu}{2}} |\mathbb{E}[\Delta_2] - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]|. \quad (20c)$$

For term (20a), applying Lemmas 2 and 3 together with the fact that the X 's and Y 's take values in some compact $\mathcal{X} \subset \mathbb{R}^d$, we obtain that

$$\sqrt{\frac{nm}{n+m}} |\text{MMD}_{\hat{K}} [X_{1:n}, Y_{1:m}] - \text{MMD}_K [X_{1:n}, Y_{1:m}]| \leq 2 \sqrt{m \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\hat{K}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})|}. \quad (21)$$

For the penultimate term in (20b), applying the bound

$$\mathbb{E} |\text{MMD}_K [X_{1:n}, Y_{1:m}] - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]| \leq 2 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right)$$

for X 's sampled from \mathbb{P} and Y 's sampled from \mathbb{Q} , whose proof can be found for instance in Section A.2 of [13], together with the bound $\sqrt{x+y} \geq (\sqrt{2}/2)(\sqrt{x} + \sqrt{y})$ for all $x, y \geq 0$, which holds due to the convexity of the square root, one has that

$$\begin{aligned} \sqrt{\frac{nm}{n+m}} |\mathbb{E}[\Delta_1] - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]| & \leq \sqrt{\frac{nm}{n+m}} \mathbb{E} |\Delta_1 - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]| \\ & \leq 2 \frac{(\sqrt{n} + \sqrt{m})}{\sqrt{n+m}} \leq \frac{4}{\sqrt{2}}. \end{aligned} \quad (22)$$

Identical arguments together with the fact that $m \leq n$ implies that $(2m)/(m+n) \leq 1$ give

$$(20b) \leq \sqrt{\frac{2m}{n+m}} \sqrt{\frac{\nu}{2}} \mathbb{E} |\Delta_2 - \text{MMD}_K [\mathbb{P}, \mathbb{Q}]| \leq \frac{4}{\sqrt{2}}. \quad (23)$$

Therefore, combining (21), (22), and (23), rearranging, and applying the rough bound

$$\mathbb{P} \left(\sum_{j=1}^K Z_j > x \right) \leq \sum_{j=1}^K \mathbb{P}(Z_j > x/K)$$

which holds for any $x \in \mathbb{R}$, $K \in \mathbb{N}$ and any random variables Z_1, \dots, Z_K , we obtain that

$$\begin{aligned} (19) & \leq \mathbb{P} \left(2 \sqrt{m \times \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\hat{K}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})|} > \right. \\ & \quad \left. \frac{1}{3} \left[\sqrt{m} \left(\frac{\sqrt{2}-1}{2\sqrt{2}} \right) \text{MMD}_K [\mathbb{P}, \mathbb{Q}] - \left(\varepsilon + \frac{10}{\sqrt{2}} \right) \right] \vee 0 \right) \\ & + \mathbb{P} \left(\sqrt{\frac{nm}{n+m}} |\Delta_1 - \mathbb{E}[\Delta_1]| > \frac{1}{3} \left[\sqrt{m} \left(\frac{\sqrt{2}-1}{2\sqrt{2}} \right) \text{MMD}_K [\mathbb{P}, \mathbb{Q}] - \left(\varepsilon + \frac{10}{\sqrt{2}} \right) \right] \vee 0 \right) \end{aligned} \quad (24a)$$

$$+ \mathbb{P} \left(\sqrt{\frac{\nu}{2}} |\Delta_2 - \mathbb{E}[\Delta_2]| > \frac{1}{3} \left[\sqrt{m} \left(\frac{\sqrt{2}-1}{2\sqrt{2}} \right) \text{MMD}_K [\mathbb{P}, \mathbb{Q}] - \left(\varepsilon + \frac{10}{\sqrt{2}} \right) \right] \vee 0 \right). \quad (24b)$$

Note that we assume $0 \leq K(\cdot, \cdot) \leq 1$. Therefore, arguing as in Lemma 4, one can show that MMD as defined in (6) has the self bounding property with constants (35). Hence, applying Theorem 5 to terms (24a) and (24b) one arrives at (19). Turning to the problem of interest we first make explicit the constants in Theorem 3:

$$\begin{aligned} C_1 &= 2 \times C_3 \\ C_2 &= \frac{1}{3} \left[\left(\frac{\sqrt{2}-1}{4} - \frac{\sqrt{50}+\sqrt{6}}{\sqrt{C_3}} \right) \right] \\ C_3 &= \left(\sqrt{6} + \sqrt{50} + \sqrt{54} \right)^2 / \left(4/(\sqrt{2}-1) \right)^2. \end{aligned}$$

Next, define the quantities

$$k^* = \min \left\{ k \in \mathbb{N} \mid k \leq \eta \text{ and } \sqrt{k} \times \text{MMD}_K[\mathbb{P}, \mathbb{Q}] \geq \sqrt{C_3 \log(2\eta/\alpha)} \right\} \quad (25a)$$

$$t_{k^*} = \min \left\{ t = 2^j \mid j \in \mathbb{N} \text{ and } t \leq k^* \right\}. \quad (25b)$$

Note that condition (11) guarantees that (25a) exists and it can be checked that $C_2 > 0$. Consequently, using (19) and the fact that $k^*/2 \leq t_{k^*} \leq k^*$, we obtain that

$$\begin{aligned} &\mathbb{P} \left((N - \eta)^+ > k^* \right) \\ &\leq \mathbb{P} \left(\bigvee_{j=0}^{\lfloor \log_2(\eta+k^*) \rfloor - 1} \sqrt{\frac{2^j(\eta+k^*-2^j)}{\eta+k^*}} \text{MMD}_{\hat{K}}[X_{A_{k^*}}, X_{B_{k^*}}] \leq \sqrt{2} + \lambda_{\eta+k^*} \right) \\ &\leq \mathbb{P} \left(\sqrt{\frac{t_{k^*}(\eta+k^*-t_{k^*})}{\eta+k^*}} \text{MMD}_{\hat{K}}[X_{1:(\eta+k^*-t_{k^*})}, X_{(\eta+k^*-t_{k^*}+1):(\eta+k^*)}] > \sqrt{2} + \lambda_{2\eta} \right) \\ &\leq \mathbb{P} \left(2\sqrt{k^* \sup_{x,y \in \mathcal{X}} |\hat{K}(x,y) - K(x,y)|} > \frac{1}{3} \left[\sqrt{k^*} \frac{\sqrt{2}-1}{4} \text{MMD}_K[\mathbb{P}, \mathbb{Q}] - \left(\lambda_{2\eta} + \frac{10}{\sqrt{2}} \right) \right] \right) \end{aligned} \quad (26a)$$

$$+ 4 \exp \left(-\frac{1}{18} \left\{ \left[\sqrt{k^*} \frac{\sqrt{2}-1}{4} \text{MMD}_K[\mathbb{P}, \mathbb{Q}] - \left(\lambda_{2\eta} + \frac{10}{\sqrt{2}} \right) \right] \vee 0 \right\}^2 \right). \quad (26b)$$

where for typographical reasons we have put:

$$\begin{aligned} A_{k^*} &:= 1 : (\eta + k^* - 2^j) \\ B_{k^*} &:= (\eta + k^* - 2^j + 1) : (\eta + k^*). \end{aligned}$$

Now (25a) together with the fact that $\lambda_{2\eta} \leq (\sqrt{50} + \sqrt{6}) \times \sqrt{\log(2\eta/\alpha)}$ for all $\alpha \in (0, 1)$ and all $\eta \in \mathbb{N}$ guarantees that the term on the right of the inequality in (26a) is no larger than $C_2 \times \sqrt{k^*} \text{MMD}_K[\mathbb{P}, \mathbb{Q}]$. Hence, appealing to (12) and Theorem 6, we obtain that (26a) $\leq \alpha/2$. Moreover, since for each $k \leq \eta$ it holds that

$$\begin{aligned} &\left\{ \sqrt{k} \left(\frac{\sqrt{2}-1}{4} \right) \text{MMD}_K[\mathbb{P}, \mathbb{Q}] - \left(\lambda_{2\eta} + \frac{10}{\sqrt{2}} \right) \geq \sqrt{18 \times 3 \log \left(\frac{2\eta}{\alpha} \right)} \right\} \\ &\subseteq \left\{ \sqrt{k} \times \text{MMD}_K[\mathbb{P}, \mathbb{Q}] \geq \sqrt{C_2 \log \left(\frac{2\eta}{\alpha} \right)} \right\} \end{aligned}$$

with k^* defined as in (25a), we obtain that (26b) $\leq 4 \times (\alpha/2\eta)^3 \leq \alpha/2$. With these facts in place the theorem is proved. \square

A.4.4 Proof of Theorem 4

Proof. Let $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^d$ and $\mathbf{0} = (0, \dots, 0)^\top \in \mathbb{R}^d$, let δ_x denote the Dirac measure (for any set $A \in \sigma(\mathbb{R}^d)$ and any $x \in \mathbb{R}^d$, $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise), and let

$$\begin{aligned} \mathcal{M}^* &= \{ \mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathbb{R}^d) \mid \mathbb{P} = p\delta_{\mathbf{1}} + (1-p)\delta_{\mathbf{0}}, \mathbb{Q} = q\delta_{\mathbf{1}} + (1-q)\delta_{\mathbf{0}}, \\ &\quad p, q \in [1/4, 3/4] \text{ and } q - p \geq 1/4 \}. \end{aligned}$$

Therefore, for any $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^*$, making use of the symmetry of K , we have that

$$\begin{aligned} & (\text{MMD}_K[\mathbb{P}, \mathbb{Q}])^2 \\ &= \mathbb{E}_{X, X' \sim \mathbb{P}} [K(X, X')] + \mathbb{E}_{Y, Y' \sim \mathbb{Q}} [K(Y, Y')] - 2\mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} [K(X, Y)] \end{aligned} \quad (27)$$

$$\begin{aligned} &= p^2 K(\mathbf{1}, \mathbf{1}) + (1-p)^2 K(\mathbf{0}, \mathbf{0}) + 2p(1-p)K(\mathbf{1}, \mathbf{0}) \\ &\quad + q^2 K(\mathbf{1}, \mathbf{1}) + (1-q)^2 K(\mathbf{0}, \mathbf{0}) + 2q(1-q)K(\mathbf{1}, \mathbf{0}) \\ &\quad - 2(pqK(\mathbf{1}, \mathbf{1}) + (1-p)(1-q)K(\mathbf{0}, \mathbf{0}) + 2(p(1-q) + q(1-p))K(\mathbf{1}, \mathbf{0})) \\ &= (K(\mathbf{1}, \mathbf{1}) + K(\mathbf{0}, \mathbf{0}) - 2K(\mathbf{1}, \mathbf{0}))(p-q)^2 \end{aligned} \quad (28)$$

Moreover, for any $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^*$, we also have that

$$\begin{aligned} \text{KL}(\mathbb{Q} \parallel \mathbb{P}) &= q \log \left(\frac{q}{p} \right) - (1-q) \log \left(\frac{1-p}{1-q} \right) \\ &\leq (q-p) \left[\frac{q}{p} - \frac{1-p}{2-(p+q)} \right] \end{aligned} \quad (29a)$$

$$\leq \frac{17}{6} (q-p) \quad (29b)$$

$$\leq \frac{34}{3} (q-p)^2, \quad (29c)$$

where (29a) holds due to the bound $\frac{x-1}{x+1} \leq \log(x) \leq x-1$ for $x \geq 1$, (29b) holds because $p, q \in [1/4, 3/4]$, and (29c) holds due to the bound $x \leq 4x^2$ for $x \geq 1/4$. Combining (28) and (29c), and additionally making use of the shift invariance of K , we obtain that

$$\text{KL}(\mathbb{Q} \parallel \mathbb{P}) \leq \frac{17}{3} (K(\mathbf{0}, \mathbf{0}) - K(\mathbf{1}, \mathbf{0}))^{-1} \text{MMD}_K^2[\mathbb{P}, \mathbb{Q}]. \quad (30)$$

Therefore, for a putting $C_K = 2(3/17) (K(\mathbf{0}, \mathbf{0}) - K(\mathbf{1}, \mathbf{0}))$ for the constant in (13), using (30) along with the fact that $\mathcal{M}^* \subset \mathcal{M}_1^+(\mathbb{R}^d)$, we have that

$$\text{L.H.S. of (13)} \geq \inf_{N: \mathbb{P}_\infty(N \leq \infty) \leq \alpha} \sup_{\substack{\eta > 1 \\ \mathbb{P}, \mathbb{Q} \in \mathcal{M}^*}} \mathbb{P} \left(N \geq \eta + \frac{(1/2) \log(1/\alpha)}{\text{KL}(\mathbb{Q} \parallel \mathbb{P})} \right). \quad (31)$$

Consequently the theorem is proved if we can find absolute constant $\alpha_0, \beta_0 \in (0, 1)$ and pre- and post-change distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^*$ such that for all $\alpha \leq \alpha_0$ it holds that

$$\inf_{N: \mathbb{P}_\infty(N \leq \infty) \leq \alpha} \sup_{\eta > 1} \mathbb{P} \left(N \geq \eta + \frac{(1/2) \log(1/\alpha)}{\text{KL}(\mathbb{Q} \parallel \mathbb{P})} \right) \geq \beta_0. \quad (32)$$

To show (32), one can use a change of measure argument originally due to Lai [27]. In fact one can directly use the version of Lai's argument adapted to finite sample analysis by Yu et al. [59, Proposition 4.1]. For clarity of exposition, we repeat the argument below. The following holds for arbitrary $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^*$. For each $n \in \mathbb{N}$ let \mathcal{F}_n be the σ -field generated by $\{X_i\}_{i=1}^n$ and let $\mathbb{P}^{\otimes n}$ be the restriction of the joint law to \mathcal{F}_n . We can write

$$\frac{d\mathbb{P}_\infty^{\otimes n}}{d\mathbb{P}_\infty^{\otimes n}} = \exp \left(\sum_{i=1}^n Z_i \right), \quad \text{for } n > \eta$$

where, as in the main text, the subscripts indicate the time at which the change occurs. For a chosen $\alpha \in (0, 1)$ and an arbitrary stopping time satisfying $\mathbb{P}_\infty(N < \infty) \leq \alpha$ introduce the events

$$\begin{aligned} \mathcal{E}_1 &= \left\{ \eta \leq N \leq \eta + \frac{(1/2) \log(1/\alpha)}{\text{KL}(\mathbb{Q} \parallel \mathbb{P})}, \sum_{i=\eta+1}^N Z_i \leq (3/4) \log(1/\alpha) \right\} \\ \mathcal{E}_2 &= \left\{ \eta \leq N \leq \eta + \frac{(1/2) \log(1/\alpha)}{\text{KL}(\mathbb{Q} \parallel \mathbb{P})}, \sum_{i=\eta+1}^N Z_i > (3/4) \log(1/\alpha) \right\}. \end{aligned}$$

For the first event we have that

$$\mathbb{P}_\eta(\mathcal{E}_1) = \int_{\mathcal{E}_1} \exp\left(\sum_{i=1}^N Z_i\right) d\mathbb{P}_\eta \leq \exp((3/4) \log(1/\alpha)) \mathbb{P}_\infty(\mathcal{E}_1) \leq \alpha^{1/4} \quad (33)$$

where the first inequality is due to the definition of \mathcal{E}_1 and the second inequality holds because the probability of N being finite when no change occurs is bounded from above by α . For the second event we have that

$$\begin{aligned} \mathbb{P}_\eta(\mathcal{E}_2) &\leq \mathbb{P}_\eta\left(\bigcup_{1=t}^{(1/2) \log(1/\alpha)(\text{KL}(\mathbb{Q} \parallel \mathbb{P}))^{-1}-1} \sum_{i=\eta+1}^{\eta+t} Z_i > (3/4) \log(1/\alpha)\right) \\ &= \mathbb{P}_\eta\left(\bigcup_{1=t}^{(1/2) \log(1/\alpha)(\text{KL}(\mathbb{Q} \parallel \mathbb{P}))^{-1}-1} \sum_{i=\eta+1}^{\eta+t} (Z_i - \text{KL}(\mathbb{Q} \parallel \mathbb{P})) > (1/4) \log(1/\alpha)\right) \end{aligned} \quad (34a)$$

$$\leq \frac{(1/2) \log(1/\alpha)}{\text{KL}(\mathbb{Q} \parallel \mathbb{P})} \exp(-\log(1/\alpha)) \quad (34b)$$

$$\leq \alpha^{1/4} \quad (34c)$$

where in particular (34a) holds by subtracting $t \times \text{KL}(\mathbb{Q} \parallel \mathbb{P})$ from both sides of the inequality and using the fact that for every t in the union it holds that $t \times \text{KL}(\mathbb{Q} \parallel \mathbb{P}) < (1/2) \log(1/\alpha)$, (34b) holds due to a union bound argument followed by an application of Hoeffding's inequality, and (34c) holds for all $\alpha \leq \alpha_0$ where

$$\alpha_0 = \sup \left\{ \alpha \in (0, 1) \mid (1/2) \log(1/\alpha) \alpha^{3/4} \leq \inf_{\mathbb{P}, \mathbb{Q} \in \mathcal{M}^*} \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \text{ and } 2\alpha^{1/4} < 1 \right\}.$$

Since the above arguments do not depend on the stopping time N or the change point location η , the bounds (33) and (34c) together imply that R.H.S. of (31) $\geq 1 - 2\alpha_0^{1/4}$, which proves the desired result. \square

A.4.5 Proof of Lemma 1

Proof. Without loss of generality assume that $m \geq n$. Let $\hat{z}_K(\cdot)$ be as defined in (8). We have that

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

Let I_1, \dots, I_n be mutually disjoint sets satisfying $\cup_{i=1}^n I_i = \{1, \dots, m\}$ and $\lfloor m/n \rfloor \leq |I_i| \leq \lfloor m/n \rfloor + 1$ for each $i = 1, \dots, n$. With $\hat{\mu}_K := \mathbb{E}_{X \sim \mathbb{P}}[\hat{z}_K(X) \mid \omega_1, \dots, \omega_r]$ introduce the quantities

$$\begin{aligned} \tilde{\eta}_i^X &= \hat{z}_K(X_i) - \hat{\mu}_K \quad \text{for } i = 1, \dots, n \\ \tilde{\eta}_j^Y &= \hat{z}_K(Y_j) - \hat{\mu}_K \quad \text{for } j = 1, \dots, m \\ \tilde{\eta}_i^{X,Y} &= \tilde{\eta}_i^X - \left\lfloor \frac{m}{n} \right\rfloor^{-1} \sum_{j \in I_i} \tilde{\eta}_j^Y \quad \text{for } i = 1, \dots, n. \end{aligned}$$

It is easy to see that $\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] = \left\| \frac{1}{n} \sum_{i=1}^n \tilde{\eta}_i^{X,Y} \right\|_2$, and that conditional on the ω 's the $\tilde{\eta}^{X,Y}$'s are independently distributed zero mean random vectors which satisfy the Bernstein condition

$$\begin{aligned} \mathbb{E} \left[\left\| \tilde{\eta}_i^{X,Y} \right\|_2^p \mid \omega_1, \dots, \omega_r \right] &\leq (2\sqrt{2})^{p-2} \mathbb{E} \left[\left\| \tilde{\eta}_i^{X,Y} \right\|_2^2 \mid \omega_1, \dots, \omega_r \right] \\ &\leq \frac{1}{2} p! (2\sqrt{2})^{p-2} \tilde{\sigma}^2(\omega_1, \dots, \omega_r), \quad \text{for } p \geq 2 \text{ and } i = 1, \dots, n, \end{aligned}$$

where, with $i' \in \arg \max_{i=1, \dots, n} |I_i|$, we let

$$\tilde{\sigma}^2(\omega_1, \dots, \omega_r) = \mathbb{E} \left[\left\| \tilde{\eta}_{i'}^{X,Y} \right\|_2^2 \mid \omega_1, \dots, \omega_r \right].$$

Therefore, applying Theorem 7, we obtain that

$$\begin{aligned}
& \mathbb{P}(\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] > \varepsilon) \\
&= \int \cdots \int \mathbb{P}(\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] > \varepsilon \mid \omega_1, \dots, \omega_r) d\Lambda(\omega_1) \cdots d\Lambda(\omega_r) \\
&\leq \int \cdots \int 2 \exp\left(-\frac{1}{2}n\varepsilon^2 \left[\tilde{\sigma}^2(\omega_1, \dots, \omega_r) + 2\sqrt{2}\varepsilon\right]^{-1}\right) d\Lambda(\omega_1) \cdots d\Lambda(\omega_r) \\
&= 2\mathbb{E}\left[\exp\left(-\frac{1}{2}n\varepsilon^2 \left[\tilde{\sigma}^2(\omega_1, \dots, \omega_r) + 2\sqrt{2}\varepsilon\right]^{-1}\right)\right] \\
&\leq 2 \exp\left(-\frac{1}{2}n\varepsilon^2 \left[\sigma^2 + 2\sqrt{2}\varepsilon\right]^{-1}\right)
\end{aligned}$$

where the final line holds due to Jensen's inequality and the fact that

$$\begin{aligned}
& \mathbb{E}[\tilde{\sigma}^2(\omega_1, \dots, \omega_r)] \\
&= \mathbb{E}[\langle \tilde{\eta}_{i'}^X, \tilde{\eta}_{i'}^X \rangle] + \left[\frac{m}{n}\right]^{-2} \sum_{j, j' \in I_{i'}} \mathbb{E}[\langle \tilde{\eta}_j^Y, \tilde{\eta}_{j'}^Y \rangle] - 2 \left[\frac{m}{n}\right]^{-1} \sum_{j \in I_{i'}} \mathbb{E}[\langle \tilde{\eta}_{i'}^X, \tilde{\eta}_j^Y \rangle] \\
&= \left\{1 + |I_{i'}| \left[\frac{m}{n}\right]^{-2}\right\} \mathbb{E}_{X \sim \mathbb{P}}[K(X, X)] \\
&\quad - \left\{2 |I_{i'}| \left[\frac{m}{n}\right]^{-1} - |I_{i'}| (|I_{i'}| - 1) \left[\frac{m}{n}\right]^{-2}\right\} \mathbb{E}_{X, Y \sim \mathbb{P}}[K(X, Y)] \\
&\leq 2\mathbb{E}_{X \sim \mathbb{P}}[K(X, X)] - \mathbb{E}_{X, Y \sim \mathbb{P}}[K(X, Y)] = \sigma^2.
\end{aligned}$$

This proves the desired result. \square

A.5 Auxiliary results

In this section, we collect a few auxiliary results. Besides establishing useful bounds on real numbers in Lemma 2 and Lemma 3, we show that the self-bounding property of RFF-MMD (Lemma 4) leads to its exponential concentration (Lemma 5). The latter is one of the key ingredients for deriving our threshold sequences elaborated Section 4.3.

Lemma 2. *For any $x, y > 0$ it holds that*

$$\frac{1}{2} \min(x, y) \leq \frac{xy}{x+y} \leq \min(x, y),$$

and, moreover, both inequalities are tight.

Proof. We first note that

$$\frac{xy}{x+y} = \frac{\min(x, y) \max(x, y)}{\min(x, y) + \max(x, y)} = \min(x, y) \left(1 + \frac{\min(x, y)}{\max(x, y)}\right)^{-1}.$$

For the lower bound we use the fact that $1 + \frac{\min(x, y)}{\max(x, y)} \leq 2$, where equality holds when $x = y$. For the upper bound, we use that $1 + \frac{\min(x, y)}{\max(x, y)} \geq 1$, where equality holds in the limit when, for instance, x is fixed and $y \rightarrow +\infty$. \square

Lemma 3. *For $x, y > 0$ it holds that $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|}$.*

Proof. When $x = y$ the statement is trivially true. When $x \neq y$ it holds that

$$|\sqrt{x} - \sqrt{y}| = \frac{|x - y|}{\sqrt{x} + \sqrt{y}} \leq \frac{|x - y|}{|\sqrt{x} - \sqrt{y}|} \Rightarrow |\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|}.$$

\square

Lemma 4. *The RFF-MMD as defined in (9) between two empirical measures composed respectively of n and m sample points is a function mapping from $(\mathbb{R}^d)^{m+n} \rightarrow \mathbb{R}$. This function has the self bounding property with constants*

$$c_i = \begin{cases} 2/n & \text{if } i = 1, \dots, n, \\ 2/m & \text{if } i = n+1, \dots, n+m. \end{cases} \quad (35)$$

Proof. Recall that if $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the reproducing kernel for some RKHS \mathcal{H}_K , for any $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+$, one has that

$$\text{MMD}_K[\mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq 1} (\mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)])$$

Note that $\hat{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ as defined in (8) is the reproducing kernel for an RKHS $\mathcal{H}_{\hat{K}}$ whose elements are vectors in \mathbb{R}^{2r} . Introduce the set

$$\hat{\mathcal{G}} = \{f \in \mathcal{H}_{\hat{K}} \mid \|f\|_{\mathcal{H}_{\hat{K}}} \leq 1\}.$$

Let $\text{MMD}_{\hat{K}}(\mathbf{x}_{1:n}, \mathbf{y}_{1:m})(\tilde{\mathbf{x}}_{i'})$ stand for (9) with inputs $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ with the i' -th \mathbf{x} replaced by $\tilde{\mathbf{x}}_{i'}$. We therefore have that

$$\begin{aligned} & \sup_{\substack{\mathbf{x}_1, \dots, \mathbf{x}_n \\ \mathbf{y}_1, \dots, \mathbf{y}_m \\ \tilde{\mathbf{x}}_{i'}}} |\text{MMD}_{\hat{K}}[\mathbf{x}_{1:n}, \mathbf{y}_{1:m}] - \text{MMD}_{\hat{K}}[\mathbf{x}_{1:n}, \mathbf{y}_{1:m}](\tilde{\mathbf{x}}_{i'})| \\ &= \sup_{\substack{\mathbf{x}_1, \dots, \mathbf{x}_n \\ \mathbf{y}_1, \dots, \mathbf{y}_m \\ \tilde{\mathbf{x}}_{i'}}} \left| \sup_{f \in \hat{\mathcal{G}}} \left(\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) - \frac{1}{m} \sum_{j=1}^m f(\mathbf{y}_j) \right) \right. \\ & \quad \left. - \sup_{f \in \hat{\mathcal{G}}} \left(\frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) - \frac{1}{n} [f(\mathbf{x}_{i'}) - f(\tilde{\mathbf{x}}_{i'})] - \frac{1}{m} \sum_{j=1}^m f(\mathbf{y}_j) \right) \right| \\ &\leq \sup_{\substack{\mathbf{x}, \tilde{\mathbf{x}} \\ f \in \hat{\mathcal{G}}}} \frac{1}{n} |f(\mathbf{x}) - f(\tilde{\mathbf{x}})| = \sup_{\mathbf{x}, \tilde{\mathbf{x}}} \frac{1}{n} \left\| \hat{K}(\cdot, \mathbf{x}) - \hat{K}(\cdot, \tilde{\mathbf{x}}) \right\|_{\mathcal{H}_{\hat{K}}} \\ &\leq \sup_{\mathbf{x}, \tilde{\mathbf{x}}} \frac{1}{n} \left(\left\| \hat{K}(\cdot, \mathbf{x}) \right\|_{\mathcal{H}_{\hat{K}}} + \left\| \hat{K}(\cdot, \tilde{\mathbf{x}}) \right\|_{\mathcal{H}_{\hat{K}}} \right) = \frac{2}{n} \end{aligned}$$

where we used the reverse triangle inequality, the reproducing property, CBS for obtaining the supremum over a unit ball and that, for any $\mathbf{x} \in \mathbb{R}^d$ one has that

$$\left\langle \hat{K}(\cdot, \mathbf{x}), \hat{K}(\cdot, \mathbf{x}) \right\rangle_{\mathcal{H}_{\hat{K}}}^2 = \cos(0) = 1.$$

The same calculations can be applied to $\text{MMD}_{\hat{K}}(\mathbf{x}_{1:n}, \mathbf{y}_{1:m})(\tilde{\mathbf{y}}_{j'})$. This proves the desired result. \square

Lemma 5. *Given two independent samples $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$, each with mutually independent entries drawn from some $\mathbb{P} \in \mathcal{M}_1^+$, for any $\varepsilon > 0$, it holds that*

$$\mathbb{P} \left(\sqrt{\frac{nm}{n+m}} \text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] > \sqrt{2} + \varepsilon \right) \leq e^{-\varepsilon^2/2}.$$

Proof. It is an immediate consequence of Lemma 4 and Theorem 5 that for any $\varepsilon' > 0$

$$\begin{aligned} & \mathbb{P} \left(\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] - \mathbb{E}[\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] \mid \omega_1, \dots, \omega_r] > \varepsilon' \mid \omega_1, \dots, \omega_r \right) \\ & \leq \exp \left(-\frac{\varepsilon'^2}{2} \frac{nm}{n+m} \right). \end{aligned} \quad (36)$$

Moreover, arguing as in the last step of the proof of Proposition 4 in [22] gives

$$\begin{aligned}
& \mathbb{E}[\text{MMD}_{\hat{K}}[X_{1:n}, Y_{1:m}] \mid \omega_1, \dots, \omega_r] \\
&= \mathbb{E} \left[\left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \hat{K}(X_i, X_j) + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \hat{K}(Y_i, Y_j) \right. \right. \\
&\quad \left. \left. - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m \hat{K}(X_i, Y_j) \right)^{\frac{1}{2}} \mid \omega_1, \dots, \omega_r \right] \\
&\leq \left(\left\{ \frac{1}{n} + \frac{1}{m} \right\} \underbrace{\mathbb{E}[\hat{K}(X, X) \mid \omega_1, \dots, \omega_r]}_{=1} \right. \\
&\quad \left. + \underbrace{\left\{ \frac{n-1}{n} + \frac{m-1}{m} - 2 \right\}}_{=-(\frac{1}{n} + \frac{1}{m})} \mathbb{E}[\hat{K}(X, Y) \mid \omega_1, \dots, \omega_r] \right)^{\frac{1}{2}} \\
&= \left(\left\{ \frac{1}{n} + \frac{1}{m} \right\} \left(1 - \underbrace{\mathbb{E}[\hat{K}(X, Y) \mid \omega_1, \dots, \omega_r]}_{\geq -1} \right) \right)^{\frac{1}{2}} \\
&\leq \sqrt{\frac{2(m+n)}{mn}}, \tag{37}
\end{aligned}$$

where the first inequality follows from Jensen's inequality. Consequently, setting $\varepsilon' = \varepsilon \sqrt{\frac{n+m}{nm}}$, plugging (37) into (36), and integrating over the ω -s with respect to the product measure $\Lambda^{\otimes r} := \Lambda \otimes \dots \otimes \Lambda$ yields the desired result. \square

A.6 External statements

In this section, we collect the external statements that we use, to ensure self-completeness. Theorem 5 recalls McDiarmid's inequality from Boucheron et al. [5, Section 6.1], which is also known as bounded differences inequality [34]. Theorem 6 is about the concentration of random Fourier features and part of the proof of Sriperumbudur and Szabó [50, Theorem 1]. We recall the concentration result Yurinsky [60, Theorem 3.3.4] on random variables taking values in a separable Hilbert space in Theorem 7.

Theorem 5 (Bounded differences inequality). *Let \mathcal{X} be a measurable space. A function $f : \mathcal{X}^n \rightarrow \mathbb{R}$ has the bounded difference property for some constants c_1, \dots, c_n if, for each $i = 1, \dots, n$,*

$$\sup_{\substack{x_1, \dots, x_n \\ x'_i \in \mathcal{X}}} |f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i. \tag{38}$$

Then, if X_1, \dots, X_n is a sequence of identically distributed random variables and (38) holds, putting $Z = f(X_1, \dots, X_n)$ and $\nu = \frac{1}{4} \sum_{i=1}^n c_i^2$ for any $t > 0$, it holds that

$$\mathbb{P}(Z - \mathbb{E}(Z) > t) \leq e^{-t^2/(2\nu)}.$$

Theorem 6 (RFF exponential concentration). *Let \hat{K} be defined as in (8). Let \mathcal{X} a proper subset of \mathbb{R}^d and denote by $|\mathcal{X}|$ its Lebesgue measure. For any $t > 0$, it holds that*

$$\mathbb{P} \left(\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \left| \hat{K}(\mathbf{x}, \mathbf{y}) - k(\mathbf{x}, \mathbf{y}) \right| > \frac{h(d, |\mathcal{X}|, \sigma) + t}{\sqrt{r}} \right) \leq e^{-t^2/2}$$

where $\sigma^2 = \int \|\omega\|_2^2 d\Lambda(\omega)$ and

$$h(d, |\mathcal{X}|, \sigma) = 23\sqrt{2d \log(2|\mathcal{X}| + 1)} + 32\sqrt{2d \log(\sigma + 1)} + 16\sqrt{2d[\log(2|\mathcal{X}| + 1)]^{-1}}. \tag{39}$$

Theorem 7 (Hilbert space Bernstein inequality). *Let X_1, \dots, X_n be a sequence of zero mean independent random variables taking values in a real and separable Hilbert space \mathcal{X} with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Write $S_n^* = \sup_{m \leq n} \|X_1 + \dots + X_m\|$. If the random variables satisfy the moment condition*

$$\mathbb{E} \|X\|^k \leq \frac{1}{2} k! B^2 H^{k-2}, \quad \text{for } k \geq 2, i = 1, \dots, n$$

for some constants $B > 0$ and $H > 0$, then for any $x > 0$ it holds that

$$\mathbb{P}(S_n^* > xB) \leq 2 \exp \left(-\frac{1}{2} x^2 \left[1 + \frac{xH}{B} \right]^{-1} \right).$$