A Contrastive Approach to Online Change Point Detection

Nikita Puchkin
HSE University and IITP RAS,
Moscow, Russian Federation

Valeriia Shcherbakova
HSE University and Sberbank of Russia,
Moscow, Russian Federation

Abstract

We suggest a novel procedure for online change point detection. Our approach expands an idea of maximizing a discrepancy measure between points from pre-change and post-change distributions. This leads to a flexible procedure suitable for both parametric and nonparametric scenarios. We prove non-asymptotic bounds on the average running length of the procedure and its expected detection delay. The efficiency of the algorithm is illustrated with numerical experiments on synthetic and real-world data sets.

1 INTRODUCTION

The problem of change point detection is familiar to statisticians and machine learners since the pioneering works of Page (1954, 1955), Shiryaev (1961, 1963) and Roberts (1966) but, nevertheless, it still attracts attention of many researchers due to its practical importance. In our paper, we assume that a learner observes independent random elements $X_1, \dots, X_t, \dots$ arriving successively. There exists a moment $\tau^* \in \mathbb{N}$ (not accessible to the statistician), such that $X_1, \dots, X_{\tau^*}$ are drawn from a distribution, which has a density p with respect to a dominating measure $\mathfrak{m}$, while $X_{\tau^*+1}, \dots, X_t, \dots$ have a density q (with respect to the same measure), which differs from p. The measure $\mathfrak{m}$ is not restricted to be the Lebesgue measure, it can be equal to the counting measure (in the discrete case) or the Hausdorff measure on a low-dimensional manifold as well. The learner is interested in reporting about the occurrence of τ^* as fast as possible while keeping the false alarm rate at an acceptable level. This problem is called online (also referred to as sequential or quickest) change point detection. Such a setup is quite different from another major research direction, offline change point detection (Dümbgen and Spokoiny, 2001; Zou et al., 2014; Matteson and James, 2014; Dalang and Shiryaev, 2015; Biau et al., 2016; Korkas and Fryzlewicz, 2017; Garreau and Arlot, 2018; Arlot et al., 2019; Madrid Padilla et al., 2021; Corradin et al., 2022; Londschien et al., 2022), where the statistician has an access to the whole time series at once, and, instead of taking decisions on the fly, he is mostly interested in a retrospective analysis and change point localization.

The complexity of a change point detection problem severely depends on the data generating mechanism. The most popular one is a mean shift, that is, $\mathbb{E}X_{\tau^*} \neq \mathbb{E}X_{\tau^*+1}$. Plenty of papers are devoted to a mean shift detection in a univariate or multivariate Gaussian sequence (see, for instance, (Enikeeva and Harchaoui, 2019; Pein et al., 2017; Rinaldo et al., 2021; Chen et al., 2022; Sun et al., 2022)), but the recent research (Eichinger and Kirch, 2018; Wang et al., 2020; Yu et al., 2020b,a) also considers a more general sub-Gaussian noise. One usually exploits CUSUM-type or likelihood-ratio-type test statistics to perform this task. A broader problem of parametric change point detection (see, for example, (Cao et al., 2018; Dette and Gösmann, 2020; Yu et al., 2020b; Corradin et al., 2022; Sun et al., 2022; Titsias et al., 2022)) admits that p and q belong to a parametric family of densities $\mathcal{P} = \{\mathsf{p}_\theta : \theta \in \Theta\}$. In this setup, the distribution change detection is reduced to detection of a shift in the underlying parameter $\theta \in \Theta$. Both the mean shift model and the parametric change point detection require strong modelling assumptions which are likely to be violated in practical applications. In our paper, we are mostly interested in a nonparametric change point detection problem (Hero, 2006; Harchaoui et al., 2008; Zou et al., 2014; Li et al., 2015; Biau et al., 2016; Garreau and Arlot, 2018; Arlot et al., 2019; Kurt et al., 2021; Shin et al., 2022). We do not impose restrictive conditions on the densities p and q. However, the procedure we propose is quite universal in a sense that it is suitable for different setups, including, for instance, the nonparametric one and the mean shift detection in a multivariate Gaussian sequence model.

Though the number of papers on change point detection is huge and many of them are devoted to theoretical analysis of the procedures (see, e.g., (Pollak and Tartakovsky, 2009; Tartakovsky et al., 2012; Li et al., 2015; Cao et al., 2018; Yu et al., 2020a; Liang et al., 2021; Chen et al., 2022; Chu and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Chen, 2022; Dehling et al., 2022; Shin et al., 2022)), nonparametric change point detection is studied not so well. Some papers provide rigorous guarantees on the average running length of the procedures (i.e. the expected number of iterations the algorithm makes in a stationary regime until a false alarm) but, to our knowledge, there are no non-asymptotic high probability bounds on the detection delay.

Let us describe the idea of our algorithm. In the sequential change point detection, at the moment t, one usually tests the hypothesis

$$H_0 : X_1, \dots, X_t \text{ have the same distribution} \tag{1}$$

against the composite alternative

$$H_1 : \text{there exists } \tau \in \{1, \dots, t-1\}, \text{ such that } \tau^* = \tau, \tag{2}$$

which can be considered as the union of the alternatives of the form $H_1^\tau : \tau^* = \tau$, $\tau \in \{1, \dots, t-1\}$. If the change occurred at some $\tau \in \{1, \dots, t-1\}$ (that is, H_1^τ takes place), then the distribution of $X_1, \dots, X_\tau$ must differ from the one of $X_{\tau+1}, \dots, X_t$. To detect such a discrepancy, we introduce an auxiliary function $D : \mathsf{X} \to (0,1)$ that should distinguish between the pre-change and post-change distributions. The higher values of $D(X)$ reflect a larger confidence that X was drawn from the density p, rather than from q. Such an approach of reducing an unsupervised learning problem to a supervised one is not new (see, e.g., (Hastie et al., 2009, Section 14.2.4)) and was used in the problems of density estimation (Gutmann and Hyvärinen, 2012), generative modelling (Goodfellow et al., 2014; Grover et al., 2019), and density ratio estimation (Grover et al., 2019). Based on this idea, Hushchyn et al. designed an algorithm for change point detection. However, the sliding window technique the authors used leads to significant detection delays. Besides, Hushchyn et al. do not provide any theoretical guarantees on the running length and the detection delay of their procedure.

Let us fix $t \in \mathbb{N}$ and a change point candidate $\tau \in \{1, \dots, t-1\}$. In order to find a good auxiliary classifier D, distinguishing between $X_1, \dots, X_\tau$ and $X_{\tau+1}, \dots, X_t$, we fix a family $\mathcal{D}$ of functions taking their values in $(0,1)$ and choose a maximizer of the cross-entropy

$$\frac{\tau(t-\tau)}{t}\left[\frac{1}{\tau}\sum_{s=1}^{\tau}\ln(2D(X_s)) + \frac{1}{t-\tau}\sum_{s=\tau+1}^{t}\ln(2-2D(X_s))\right] \tag{3}$$

over $\mathcal{D}$. A similar approach was introduced in (Gutmann and Hyvärinen, 2012; Goodfellow et al., 2014) but for the purposes of density estimation and generative modelling, respectively. In the context of sequential change point detection, Li et al. and Chang et al. used a different divergence measure, the squared maximum mean discrepancy, to derive a kernel change point detection method. In our paper, we adapt the technique of (Goodfellow et al., 2014) for the quickest change point detection. Following (Gutmann and Hyvärinen, 2012; Goodfellow et al., 2014), we call our approach *contrastive* and refer to the function D as discriminator.

We show in Section 2.1 that our algorithm needs to approximate $\ln(\mathsf{p}/\mathsf{q})$ with a reasonable accuracy to be sensitive to distribution changes. This makes it similar to change point detection methods based on the density ratio estimation (Liu et al., 2013; Hushchyn et al., 2020; Hushchyn and Ustyuzhanin, 2021). For instance, Liu et al. uses KLIEP (Sugiyama et al., 2008), uLSIF (Kanamori et al., 2009) and RuLSIF (Yamada et al., 2013) for online change point detection. In (Hushchyn and Ustyuzhanin, 2021), the authors use the α-relative chi-squared divergence, the same functional as in RuLSIF (Yamada et al., 2013), to construct a change point detection procedure. The advantage of such methods is that the estimation of the ratio p/q can be a much easier task than estimation of the densities p and q themselves. However, in the density-ratio based algorithms the authors usually use a sliding window technique and compare the distributions between two large non-overlapping segments of the time series. This approach shows a good performance in the offline setup, when the learner is interested in change point estimation, but leads to large detection delays in the online case. In our paper, we adjust the test statistic in order to make it suitable for the sequential detection problem. Besides, in contrast to (Liu et al., 2013; Hushchyn et al., 2020; Hushchyn and Ustyuzhanin, 2021), we study the detection delay of our procedure and the behaviour of the test statistic under the null hypothesis.

Contribution. We suggest a procedure for sequential change point detection based on the contrastive approach. We provide non-asymptotic large deviation bounds on the running length and the detection delay of our procedure (Theorems 2.7 and 2.9) for general classes of discriminators. We also specify the results of these theorems for particular cases, including nonparametric change point detection via neural networks. To our knowledge, Corollary 3.1 is the first theoretical guarantee for such a setup. Finally, we illustrate the performance of our procedure with numerical experiments on synthetic and real-world data sets.

Organization of the paper. The rest of the paper is organized as follows. In Section 2, we introduce our algorithm and discuss its theoretical properties. In particular, we derive non-asymptotic large deviation bounds on the running length and the detection delay of our procedure (Theorems 2.7 and 2.9). In Section 3, we specify the result of Theorem 2.7 for the case when $\ln(\mathsf{p}/\mathsf{q})$ is a smooth function. We also show how the result of Theorem 2.9 yields an almost optimal mean shift detection in a Gaussian sequence model. Section 4 is devoted to numerical experiments. Proofs of the theoretical results are deferred to the supplemental ma-

terial.

Notation. We use the following notations throughout the paper. The notation $f \lesssim g$ or $g \gtrsim f$ means that $f \leqslant cg$ for an absolute constant c. We also use the standard $O(\cdot)$ notation. To avoid problems with the logarithmic function, we use the convention $\log x = (1 \vee \ln x)$. We set $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$, and $a_+ = \max\{a, 0\}$. For $s \geqslant 1$ and a measure with the density p, we define the $L_s(\mathsf{p})$-norm as $\|f\|_{L_s(\mathsf{p})} = (\mathbb{E}_{\xi\sim\mathsf{p}}|f(\xi)|^s)^{1/s}$, $L_\infty(\mathsf{p})$-norm as $\|f\|_{L_\infty(\mathsf{p})} = \operatorname{esssup}|f(\xi)|$, where $\xi \sim \mathsf{p}$, and the $\psi_s(\mathsf{p})$-norm as $\|f\|_{\psi_s(\mathsf{p})} = \inf\{u > 0 : \mathbb{E}_{\xi\sim\mathsf{p}} \exp(|f(\xi)|^s/u^s) \leqslant 2\}$. We use the notation $\psi_s(\mathsf{p})$ for the Orlicz norm, instead of the conventional ψ_s, to specify a probability measure and avoid ambiguity, since we deal with different pre-change and post-change distributions. For a class of functions $\mathcal{F}$, equipped with a norm $\|\cdot\|$, we denote its diameter (with respect to $\|\cdot\|$) by $\mathcal{D}(\mathcal{F}, \|\cdot\|)$. Given two probability measures with the densities $\mathsf{p} \ll \mathsf{q}$, $\mathrm{KL}(\mathsf{p}, \mathsf{q}) = \int \mathsf{p}(x) \ln(\mathsf{p}(x)/\mathsf{q}(x)) \mathrm{d}\mathfrak{m}$ stands for the Kullback-Leibler divergence between p and q. For any two densities p and q, $\mathrm{JS}(\mathsf{p}, \mathsf{q}) = \mathrm{KL}(\mathsf{p}, (\mathsf{p} + \mathsf{q})/2)/2 + \mathrm{KL}(\mathsf{q}, (\mathsf{p}+\mathsf{q})/2)/2$ denotes the Jensen-Shannon divergence between p and q.

2 THE ALGORITHM AND ITS THEORETICAL PROPERTIES

In this section, we present our procedure, given in Algorithm 1, and then discuss its theoretical properties. On each iteration $t \in \mathbb{N}$ and for each change point candidate $\tau \in \{1, \ldots, t-1\}$, the algorithm tries to maximize the discrepancy measure (3). The requirement that the classifier D in (3) must take its values in $(0, 1)$ is inconvenient in practical tasks. To avoid this issue, we use the standard reparametrization $D(x) = e^{f(x)}/(1 + e^{f(x)})$, $f \in \mathcal{F}$, obtain the functional $\mathcal{T}_{\tau,t}(f)$ of the form (4) and find its maximizer $\widehat{f}_{\tau,t}$. After that, we compute a test statistic $\mathcal{S}_t$ as the maximum of $\mathcal{T}_{\tau,t}(\widehat{f}_{\tau,t})$ with respect to τ.

At the round t, Algorithm 1 solves $t-1$ optimization problems. If the class $\mathcal{F}$ is convex, then, using the standard gradient ascent, one can find an ε-maximizer of the functional $\mathcal{T}_{\tau,t}(f)$ in just $O(\log(1/\varepsilon))$ iterations, because of the strong convexity of $\mathcal{T}_{\tau,t}$. Unfortunately, it requires $O(t)$ operations to compute the gradient of $\mathcal{T}_{\tau,t}(f)$. As an alternative, one may use the stochastic gradient ascent to reduce the computational cost of the gradient to $O(1)$. However, such an improvement is not for free, since the stochastic gradient algorithm requires $O(1/\varepsilon)$ iterations to get an ε-maximizer. To sum up, for any $\varepsilon \in (0, 1)$, the procedure requires $O(t^2 \log(1/\varepsilon) \wedge t/\varepsilon)$ operations to compute $\mathcal{S}_t$ within the accuracy ε. This may become prohibitive with the growth of t, and we suggest restarting the procedure from time to time. The good news is that the changes one needs to detect are quite steep in many real-life scenarios, so one does not have to take ε too small nor the class $\mathcal{F}$ too broad. We show in Section 4 that it is enough to take a class $\mathcal{F}$ of simple structure for consistent change point detection.

Algorithm 1 Contrastive online change point detection

Require: a class of functions $\mathcal{F}$ and a threshold $\mathfrak{z} > 0$.

1: **for** $t = 1, 2, \ldots$ **do** the following

2: Receive an observation X_t.

3: For each $\tau \in \{1, \ldots, t-1\}$, compute the estimates $\widehat{f}_{\tau,t} \in \operatorname*{argmax}_{f\in\mathcal{F}} \mathcal{T}_{\tau,t}(f)$, where

$$\mathcal{T}_{\tau,t}(f) = \frac{t-\tau}{t} \sum_{s=1}^{\tau} \left[f(X_s) - \ln\left(\frac{1 + e^{f(X_s)}}{2}\right)\right] - \frac{\tau}{t} \sum_{s=\tau+1}^{t} \ln\left(\frac{1 + e^{f(X_s)}}{2}\right). \quad (4)$$

4: Compute the test statistic

$$\mathcal{S}_t = \max_{1\leqslant\tau\leqslant t-1} \mathcal{T}_{\tau,t}(\widehat{f}_{\tau,t}). \quad (5)$$

5: If $\mathcal{S}_t > \mathfrak{z}$, terminate the procedure, and report the change point occurrence.

return

2.1 Behaviour of the test statistic in the presence of a change point

We start with an analysis of the behaviour of the statistic $\mathcal{T}_{\tau,t}(f)$ in the presence of a change point.

Lemma 2.1. *Fix $t \in \mathbb{N}$ and let $f^*(x) = \ln(\mathsf{p}(x)/\mathsf{q}(x))$. Assume that the change point occurred at some $\tau \in \{1, 2, \ldots, t-1\}$, that is, $\tau^* = \tau$. Then, for any measurable function f, it holds that*

$$\mathbb{E}\mathcal{T}_{\tau,t}(f) \geqslant \frac{2\tau(t-\tau)}{t} \left(\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \frac{1}{16}\|f - f^*\|^2_{L_2(\mathsf{p}+\mathsf{q})} \right). \quad (6)$$

where $\mathrm{JS}(\mathsf{p}, \mathsf{q})$ is the Jensen-Shannon divergence between p and q.

Lemma 2.1 illustrates two important properties of $\mathcal{T}_{\tau,t}(f)$. First, if a change point occurred, then, for any $f \in \mathcal{F}$ the expectation of $\mathcal{T}_{\tau^*,t}(f)$ (and, as consequence, the expectation of $\mathcal{S}_t$) grows as the detection delay $(t - \tau^*)$ increases. We show in the proofs of Theorems 2.7 and 2.9 that the actual value of $\mathcal{S}_t$ will not be much smaller than its expectation with high probability due to the concentration of measure phenomenon. On the other hand, Lemma 2.1 reveals a relation of our procedure with change point detection methods based on density ratio estimation. As one can conclude from (6), the class $\mathcal{F}$ must be chosen in a way to approximate $\ln(\mathsf{p}(x)/\mathsf{q}(x))$ with a reasonable accuracy. At the same time, as a reader will see in the next section,

a broader class $\mathcal{F}$ yields larger values of the test statistics under the null hypothesis. A practitioner must keep this trade-off in mind while choosing $\mathcal{F}$.

2.2 Behaviour of the test statistic under the null hypothesis

In this section, we study the behaviour of the test statistic $\mathcal{T}_{\tau,t}(\widehat{f})$ in two scenarios. The first one, considered in Theorem 2.3, concerns the case when $\mathcal{F}$ is a class of functions taking their values in $[-B, B]$ for some $B > 0$. A possible extension for unbounded classes is discussed in Theorem 2.6.

Before we formulate the theoretical results rigorously, let us remind a reader some preliminaries on covering and bracketing numbers. Given a normed space $(\mathcal{F}, \|\cdot\|_{L_2(\mathsf{p})})$ and $u > 0$, the covering number $\mathcal{N}(\mathcal{F}, L_2(\mathsf{p}), u)$ is the minimal number of balls of radius u needed to cover $\mathcal{F}$. Further, for any $f_1, f_2 \in \mathcal{F}$, such that $f_1 \leqslant f_2$ almost surely, a bracket $[f_1, f_2]$ is a set of all such $g \in \mathcal{F}$ that $f_1 \leqslant g \leqslant f_2$ with probability one. The size of the bracket $[f_1, f_2]$ is $\|f_1 - f_2\|_{L_2(\mathsf{p})}$. The bracketing number $\mathcal{N}_{[]}(\mathcal{F}, L_2(\mathsf{p}), u)$ is the minimal number of brackets of size u needed to cover $\mathcal{F}$.

In the bounded case, we require the class $\mathcal{F}$ to have a polynomial bracketing number. In Section 3, we give an example that, if p is supported on a unit cube in $\mathbb{R}^p$, then a class of neural networks with ReLU activations satisfies this assumption.

Assumption 2.2. *There exist positive constants A, B, and d, such that $\mathcal{D}(\mathcal{F}, L_\infty(\mathsf{p})) \leqslant B$, and the bracketing number of the class $\mathcal{F}$ with respect to the $L_2(\mathsf{p})$-norm satisfies the inequality*

$$\mathcal{N}_{[]}(\mathcal{F}, L_2(\mathsf{p}), u) \leqslant \left(\frac{A}{u}\right)^d, \quad \text{for all } 0 < u \leqslant 2B.$$

We are in position to formulate a result about the large deviations of $\mathcal{T}_{\tau,t}(\widehat{f}_{\tau,t})$.

Theorem 2.3. *Grant Assumption 2.2. Fix any $t \in \mathbb{N}$, any $\tau \in \{1, \dots, t-1\}$ and assume that $X_1, \dots X_t$ are i.i.d. random elements with the density p. Let $\widehat{f} \in \operatorname{argmax}_{f\in\mathcal{F}} \mathcal{T}_{\tau,t}(f)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\mathcal{T}_{\tau,t}(\widehat{f}) \lesssim de^B \left[B + \log\left(\frac{A\tau(t-\tau)}{td}\right)\right] + e^B \log(1/\delta). \tag{7}$$

Note that if $0 \in \mathcal{F}$ (which is a very mild requirement), then the statistic $\mathcal{T}_{\tau,t}(\widehat{f})$ is non-negative. Theorem 2.3 also shows that we use a proper scaling for $\mathcal{T}_{\tau,t}(\widehat{f})$ in a sense that the high probability upper bound for $\mathcal{T}_{\tau,t}(\widehat{f})$ has only a logarithmic dependence on τ.

Unfortunately, the boundedness of $\mathcal{F}$ with respect to the $L_\infty(\mathsf{p})$-norm may be restrictive, especially if p has an unbounded support. In the rest of this section, we consider the case when $\mathcal{D}(\mathcal{F}, L_\infty(\mathsf{p}))$ is allowed to be infinite.

Definition 2.4. *A class of functions $\mathcal{F}$ is called L-sub-Gaussian (with respect to a density p) if $\|f\|_{\psi_2(\mathsf{p})} \leqslant L\|f\|_{L_2(\mathsf{p})}$ for all $f \in \mathcal{F}$.*

A simple example of a sub-Gaussian class is the class of linear functions (with respect to a Gaussian measure). In Section 3, we show that Algorithm 1 with the linear class $\mathcal{F}$ can efficiently detect a mean shift in a multivariate Gaussian sequence model. We also relax the bracketing number assumption and replace it by the next one.

Assumption 2.5. *The class $\mathcal{F}$ is L-sub-Gaussian for some constant $L > 0$. Besides, there exist positive constants A and d, such that the covering number of the class $\mathcal{F}$ with respect to the $L_2(\mathsf{p})$-norm satisfies the inequality*

$$\mathcal{N}(\mathcal{F}, L_2(\mathsf{p}), u) \leqslant \left(\tfrac{A}{u}\right)^d, \quad \text{for all } 0 < u \leqslant \mathcal{D}(\mathcal{F}, L_2(\mathsf{p})).$$

We are ready to formulate our main result concerning the behaviour of the statistic $\mathcal{T}_{\tau,t}(\widehat{f}_{\tau,t})$ in the stationary regime in the unbounded case.

Theorem 2.6. *Grant Assumption 2.5. Fix any $t \in \mathbb{N}$, any $\tau \in \{1, \dots, t-1\}$ and assume that $X_1, \dots X_t$ are i.i.d. random elements with the density p. Let $\widehat{f} \in \operatorname{argmax}_{f\in\mathcal{F}} \mathcal{T}_{\tau,t}(f)$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\mathcal{T}_{\tau,t}(\widehat{f}) \lesssim L^2 d e^{\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{2\log(4L\sqrt{2})}} \log(1/\delta) \cdot \left[\mathcal{D}(\mathcal{F}, \psi_2(\mathsf{p}))\sqrt{\log L} + \log\left(\frac{A\tau(t-\tau)}{L^2 td}\right)\right].$$

To sum up the results of Sections 2.1 and 2.2, the test statistic S_t, defined in (5), is expected to grow steeply after the change point. On the other hand, it will be not too large in the stationary regime, when no change point occurs. Let us illustrate this point with a simple example. Let $X_1, \dots, X_T$, $T = 100$, be a sequence of i.i.d. observations drawn according to the Gaussian distribution $\mathcal{N}(0, 0.01)$. Let $\tau^* = 75$ and define a sequence $Y_1, \dots, Y_T$ according to the formula

$$Y_t = \begin{cases} X_t, & \text{if } t \leqslant \tau^*, \\ 0.2 + X_t, & \text{otherwise.} \end{cases}$$

In other words, the sequences $\{X_t : 1 \leqslant t \leqslant T\}$ and $\{Y_t : 1 \leqslant t \leqslant T\}$ coincide before the change point τ^* and differ by the shift $\mu = 0.2$ after it. A realization of the sequences is displayed in Figure 1. Let $\mathcal{F} = \{f(x) = wx + b : w, b \in \mathbb{R}\}$ be a class of linear functions and apply Algorithm 1 to the sequences described above. We observe that the test statistic $\mathcal{S}_t$, computed for the sequence $Y_1, \dots, Y_T$ (the solid red line in Figure 1), rises

sharply after the change point (see Figure 1, bottom line) and achieves the value 17.5 in the end but, on the other hand, it does not exceed 2.5 in the stationary regime (see the dotted blue line in Figure 1).

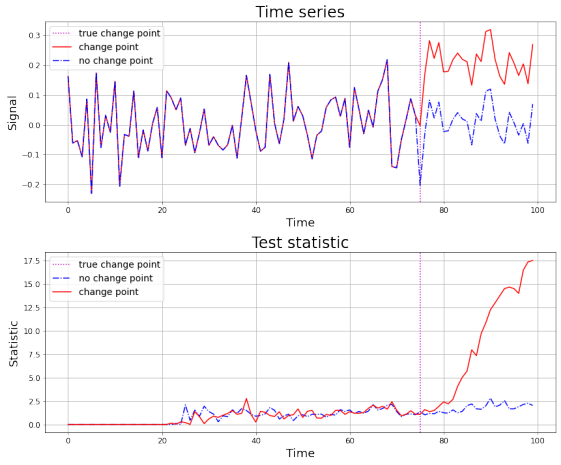


Figure 1: An example of behaviour of the test statistic $\mathcal{S}_t$ (defined in (5)) in the presence of a change point and in the stationary regime. Top line: a stationary sequence (blue) and a sequence of observations with a change point (red). Bottom line: corresponding values of the test statistic $\mathcal{S}_t$ with a linear class. The dashed vertical line corresponds to the true change point τ^*.

2.3 Bounds on the average running length and the expected detection delay

In this section, we provide lower bounds on the average running length and the expected detection delay of Algorithm 1, based on our findings presented in Sections 2.1 and 2.2. All the bounds hold with high probability.

Let us introduce $\rho(\mathcal{F}) = \inf_{f\in\mathcal{F}} \|\ln(\mathsf{p}/\mathsf{q}) - f\|_{L_2(\mathsf{p}+\mathsf{q})}$. We have the following result for the case of $\mathcal{F}$ with bounded diameter with respect to the $L_\infty(\mathsf{p})$-norm.

Theorem 2.7. *Grant Assumption 2.2 and assume that $\|f\|_{L_\infty(\mathsf{q})} \leqslant B$ for all $f \in \mathcal{F}$. Fix any $\delta \in (0,1)$ and $T \in \mathbb{N}$. There exists such a choice of $\mathfrak{z}$ in Algorithm 1 (specified in the proof) that the following holds:*

- *if $\tau^* = \infty$, then Algorithm 1 makes at least T steps until the false alarm with probability at least $1-\delta$;*
- *otherwise, if τ^* is sufficiently large, so that it fulfils*

$$\tau^* \geqslant \frac{B^2\log(1/\delta)}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{6B\log(1/\delta) + 2\mathfrak{z}}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)},$$

then the stopping time $\widehat{t}$ of Algorithm 1 satisfies the inequality

$$\widehat{t} - \tau^* \lesssim \frac{B^2\log(1/\delta)}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{de^B\left[B + \log\left(AT/d\right)\right] + e^B\log(T/\delta)}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)}$$

with probability at least $1-\delta$.

Remark 2.8. *We emphasize that $\widehat{t}$ is the stopping time of the procedure, it should not be confused with an estimate of τ^*. A natural way to estimate the change point is to consider*

$$\widehat{\tau} = \operatorname*{argmax}_{\tau\in\{1,\dots,\widehat{t}\}} S_{\tau,\widehat{t}}.$$

However, in the present paper, we focus on the running length and the detection delay only. We do not tackle the problem of estimation of τ^*. The study of theoretical properties of $\widehat{\tau}$ is left for the future research.

Using a similar technique as in the proof of Theorem 2.7 and Theorem 2.6, we obtain large deviation bounds on the detection delay of Algorithm 1 in the sub-Gaussian case.

Theorem 2.9. *Grant Assumption 2.5 and fix any $\delta \in (0,1)$, $T \in \mathbb{N}$. There exists such a choice of $\mathfrak{z}$ in Algorithm 1 (specified in the proof) that the following holds:*

- *if $\tau^* = \infty$, then Algorithm 1 makes at least T steps until the false alarm with probability at least $1-\delta$;*
- *otherwise, if τ^* is sufficiently large, then the stopping time $\widehat{t}$ of Algorithm 1 satisfies the inequality*

$$\widehat{t} - \tau^* \lesssim \frac{L^2 d\log(T/\delta)e^{\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{2\log(4L\sqrt{2})}}}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)} \cdot \left[\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{\log L} + \log\left(\frac{AT}{Ld}\right)\right] + \frac{\left[\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F},\psi_2(\mathsf{q}))\right]^2\log(1/\delta)}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)^2}$$

with probability at least $1-\delta$.

Remark 2.10. *The results of Theorems 2.7 and 2.9 can be easily extended to the case of multiple change points. We just have to restart the procedure after a structural break was detected. The only additional requirement will be that the distance between two subsequent change points is $\Omega(\log T)$, which is quite standard for the offline setup.*

We elaborate on the results of Theorems 2.7 and 2.9 in the next section.

3 EXAMPLES

In this section, we specify the results of Theorem 2.7 and Theorem 2.9 for particular cases. Our examples include nonparametric change point detection via feed-forward neural networks and the classical problem of mean shift detection in a Gaussian sequence model.

3.1 Nonparametric online change point detection via neural networks

Consider the following nonparametric change point detection setup. Let $X_1, \dots, X_t, \dots$ be independent random elements and assume that $X_1, \dots, X_{\tau^*}$ have a density p supported on $[0,1]^p$ while the other elements of the time series are drawn from the density q (also supported on the unit cube $[0,1]^p$). Assume that $\ln(\mathsf{p}/\mathsf{q})$ belongs to a Hölder class $\mathcal{H}^\beta([0,1]^p, H)$ for some smoothness parameter $\beta > 0$ and some $H > 0$. Recall that the class $\mathcal{H}^\beta([0,1]^p, H)$ is defined as

$$\mathcal{H}^\beta([0,1]^p, H) = \Bigg\{ f : [0,1]^p \to \mathbb{R} : \sum_{\substack{\alpha \in \mathbb{Z}_+^p : \\ \|\alpha\|_1 < \beta}} \|\partial^\alpha f\|_{L_\infty([0,1]^p)} + \sum_{\substack{\alpha \in \mathbb{Z}_+^p : \\ \|\alpha\|_1 = \lfloor \beta \rfloor}} \sup_{\substack{x,y \in [0,1]^p \\ x \neq y}} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leqslant H \Bigg\},$$

where, for any multi-index $\alpha = (\alpha_1, \dots, \alpha_p)$, $\partial^\alpha f(x)$ stands for the partial derivative $\partial^{\alpha_1} \dots \partial^{\alpha_p} f(x) / (\partial x_1^{\alpha_1} \dots \partial x_p^{\alpha_p})$ and, for any $\beta \in \mathbb{R}$, $\lfloor \beta \rfloor$ denotes the largest integer strictly less than β. We use a class of feed-forward neural networks to detect a change in the distribution of the observed sequence. Introduce the ReLU activation function $\sigma(u) = (u \vee 0)$, $u \in \mathbb{R}$, and, for any $v \in \mathbb{R}^p$, define the *shifted* activation function σ_v as $\sigma_v(x) = (\sigma(x_1 - v_1), \dots, \sigma(x_p - v_p))^\top$, $x \in \mathbb{R}^p$. A neural network is a composition of linear and nonlinear maps. Let us fix a number of hidden layers $L \in \mathbb{N}$, an architecture $\mathcal{A} = (a_0, \dots, a_{L+1}) \in \mathbb{N}^{L+2}$, matrices $W_1 \in \mathbb{R}^{a_1 \times a_0}, \dots, W_{L+1} \in \mathbb{R}^{a_{L+1} \times a_L}$, and shift vectors $v_1 \in \mathbb{R}^{a_1}, \dots, v_L \in \mathbb{R}^{a_L}$. Then a neural network with ReLU activations, L hidden layers and the architecture $\mathcal{A}$ is a function $f : \mathbb{R}^{a_0} \to \mathbb{R}^{a_{L+1}}$ of the form

$$f = W_L \circ \sigma_{v_L} \circ W_{L-1} \circ \sigma_{v_{L-1}} \circ \dots \circ W_1 \circ \sigma_{v_1} \circ W_0. \quad (8)$$

Given $L \in \mathbb{N}, \mathcal{A} = (p, a_1, \dots, a_L, 1) \in \mathbb{N}^{L+2}$, and $s \in \mathbb{N}$, we consider a class of sparsely connected neural networks

$$\mathsf{NN}(L, \mathcal{A}, s) = \Bigg\{ f \text{ of the form (8)} : \max_{i,j} |W_{ij}| \vee \max_i |v_i| \leqslant 1, \ \|W_0\|_0 + \sum_{j=1}^{L} (\|W_j\|_0 + \|v_j\|_0) \leqslant s \Bigg\}.$$

To our knowledge, this class of neural networks was first studied in (Schmidt-Hieber, 2020). The sparsity parameter s is introduced to reflect the fact that, in practice, one rarely uses fully connected neural networks, and the number of active neurons is usually much smaller than the total number of parameters $\sum_{j=1}^{L} a_j + \sum_{j=1}^{L+1} a_{j-1} a_j$. In (Schmidt-Hieber, 2020), the author proves two important results, concerning approximation properties (Theorem 5) and the covering number of the class $\mathsf{NN}(L, \mathcal{A}, s)$ (Lemma 5). We provide their statements in Appendix B to make the paper self-contained. Combining these results with Theorem 2.7, we get the following corollary.

Corollary 3.1. *Assume that $\ln(\mathsf{p}/\mathsf{q}) \in \mathcal{H}^\beta([0,1]^p, H)$. Fix any $\delta \in (0,1)$ and $T \in \mathbb{N}$. There exist $C > 0$, $\tau_\circ \in \mathbb{N}$, $L \in \mathbb{Z}_+, \mathcal{A} \in \mathbb{N}^{L+2}$, $s \in \mathbb{N}$, and $\mathfrak{z} > 0$ (specified in the proof) such that the following holds. Run Algorithm 1 with the class of truncated neural networks*

$$\mathsf{NN}_B(L, \mathcal{A}, s) = \big\{ g(x) = -B \vee (f(x) \wedge B) : f \in \mathsf{NN}(L, \mathcal{A}, s) \big\}, \quad (9)$$

where B any number greater than $H + \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}$. If $\tau^* = \infty$, then Algorithm 1 stops after at least T steps with probability at least $1 - \delta$. Otherwise, if τ^* is sufficiently large, then, with probability at least $1 - \delta$, the stopping time $\widehat{t}$ of Algorithm 1 fulfils

$$\widehat{t} - \tau^* \lesssim \frac{e^B \log(1/\mathrm{JS}(\mathsf{p}, \mathsf{q})) \, [B + \log(1/\mathrm{JS}(\mathsf{p}, \mathsf{q})) \log T]}{\mathrm{JS}(\mathsf{p}, \mathsf{q})^{\frac{2\beta + p}{2\beta}}} + e^B \log(T/\delta) + \frac{B^2 \log(1/\delta)}{\mathrm{JS}(\mathsf{p}, \mathsf{q})^2},$$

where the hidden constant depends on H, β, and p.

To our knowledge, this is the first non-asymptotic high probability bound on the detection delay for an online change point detection procedure exploiting neural networks.

3.2 Online detection of a mean shift in a Gaussian sequence model

In this section, we show how the result of Theorem 2.9 applies to the classical problem of mean shift detection in a Gaussian sequence. Assume that $X_1, \dots, X_{\tau^*}$ are i.i.d. random vectors in $\mathbb{R}^p$ with the Gaussian distribution $\mathcal{N}(0, \Sigma)$ while $X_{\tau^*+1}, \dots, X_t, \dots$ have the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, $\mu \neq 0$. In this case, $\ln(\mathsf{p}/\mathsf{q})$ is linear, so it is reasonable to consider the following class of functions:

$$\mathcal{F}_{\text{lin}} = \big\{ f_{w,b}(x) = w^\top x + b : \|\Sigma^{1/2} w\| \leqslant \|\Sigma^{-1/2} \mu\|, |b| \leqslant \mu^\top \Sigma^{-1} \mu \big\}. \quad (10)$$

With this choice, $f^* \in \mathcal{F}_{\text{lin}}$, that is, the approximation error is equal to zero. At the same time, the class $\mathcal{F}_{\text{lin}}$ is sub-Gaussian, its ψ_2-diameter is of order $(\mu^\top \Sigma^{-1} \mu)^{1/2}$, and its metric entropy satisfies the inequality $\log \mathcal{N}(\mathcal{F}_{\text{lin}}, L_2(\mathsf{p}), \varepsilon) \lesssim p \log(\mu^\top \Sigma^{-1} \mu / \varepsilon)$ for any $\varepsilon > 0$. Substituting these bounds into the statement of Theorem 2.9, we get the following corollary.

Corollary 3.2. *Assume that $\|\Sigma^{-1/2} \mu\| \leqslant \ln(4/3)$. Fix any $\delta \in (0,1)$ and $T \in \mathbb{N}$. There exists such $\mathfrak{z} > 0$ that the following holds:*

- *if $\tau^* = \infty$, then the running length of Algorithm 1 (with the class $\mathcal{F}_{lin}$ given by (10)) is at least T on an event with probability $1 - \delta$;*
- *otherwise, if τ^* is sufficiently large, then the stopping time $\widehat{t}$ of Algorithm 1 (with the class $\mathcal{F}_{lin}$) satisfies the inequality*

$$\widehat{t} - \tau^* \lesssim \frac{p \log\big(\|\Sigma^{-1/2}\mu\| T/p\big) \log(T/\delta)}{\mu^\top \Sigma^{-1} \mu}$$

with probability at least $1 - \delta$.

The number $\|\Sigma^{-1/2}\mu\|$ is sometimes referred to as signal-to-noise ratio (SNR) in the change point detection literature. For the ease of presentation, we consider only the case of low SNR. Corollary 3.2 shows that our approach captures the dependence on $\|\Sigma^{-1/2}\mu\|$ correctly (cf. (Yu et al., 2020a, Theorem 1) and (Chen et al., 2022, Theorem 2)). However,it has an additional logarithmic term compared to the worst-case detection delay of the CUSUM-type procedure (Yu et al., 2020a). This artefact appears because of universality of our algorithm: it tries to learn a an optimal discriminator D without any prior knowledge about p and q nor about the kind of change, while the CUSUM-type procedure (Yu et al., 2020a) exploits the difference in means of the pre-change and post-change distributions.

4 NUMERICAL EXPERIMENTS

In this section, we illustrate the performance of our procedure on synthetic and real-world data sets. The code for all the experiments, described below, is available at Github[1]. We consider three variants of the class $\mathcal{F}$ used in Algorithm 1. The first one is the class of polynomials of degree p. The second is the linear span of the first q elements of the Fourier basis $1, \sin(2\pi x), \cos(2\pi x), \sin(4\pi x)$, etc. Finally, the third class is the class of fully connected feed-forward neural networks with architecture $(1, 2, 3, 1)$ and ReLU activations. The neural network architecture was the same in all the experiments. We truncate the function values if their absolute value exceeds 10 to avoid numerical issues. Though the class $\mathcal{F}$ 1 can be arbitrary in general, Theorems 2.7 and 2.9 imply that it should be expressive enough to approximate the log-ratio $\ln(\mathsf{p}/\mathsf{q})$ with an adequate accuracy. Our choice is motivated by the fact that polynomials, trigonometric Fourier series and feed-forward neural networks with ReLU activations are known to have the universal approximation property. One can consider other variants of the class $\mathcal{F}$.

The performance of our method (with three aforementioned variants of the class $\mathcal{F}$) is compared with two popular nonparametric change point detection methods: KLIEP (Sugiyama et al., 2008; Liu et al., 2013) and kernel change point detection with M-statistic (Li et al., 2015). We also added the comparison with CUSUM (see, e.g., (Wang et al., 2020, Definition 1)) in the experiment with a shift in expectation. KLIEP is a density-ratio-based change point detection method, estimating of the KL-divergence between the pre-change and post-change distributions. As we discussed in Section 2.1, our approach is somehow related to density-ratio-based methods, so it is reasonable to compare our algorithm with one of them. In Li et al. (2015), the authors use kernel methods to approximate the squared maximum mean discrepancy (MMD) between the pre-change and post change distributions. We use a different divergence measure, based on the maximum cross-entropy, but the core idea of maximizing discrepancy between pre-change and post change observations is quite similar.

4.1 Synthetic data sets

The experiments with synthetic data check the ability of the procedure to detect changes in mean, variance, and the density of the distribution. Before we move to description of our results, let us elaborate on how we tune the thresholds. We sample $T = 150$ i.i.d. samples according to p and compute the maximal value of the corresponding test statistic $\mathcal{S}_t^{(1)}$, $1 \leqslant t \leqslant 150$. We repeat the procedure several times and obtain the values $\max_{1 \leqslant y \leqslant T} \mathcal{S}_t^{(2)}, \ldots, \max_{1 \leqslant y \leqslant T} \mathcal{S}_t^{(J)}$, where J is the number of repetitions. Then we put $\mathfrak{z} = \max_{1 \leqslant j \leqslant J} \max_{1 \leqslant t \leqslant T} \mathcal{S}_t^{(j)}$. Such a choice ensures that the running length of our procedure is not smaller than $T = 150$ with probability at least $1 - 1/(J+1)$. Indeed, if we run the procedure in the stationary regime and compute the corresponding values of the test statistic $\mathcal{S}_t$, then the probability that $\max_{1 \leqslant t \leqslant T} \mathcal{S}_t$ exceeds $\mathfrak{z} = \max_{1 \leqslant j \leqslant J} \max_{1 \leqslant t \leqslant T} \mathcal{S}_t^{(j)}$ is the same as $\max_{1 \leqslant t \leqslant T} \mathcal{S}_t^{(j)}$ exceeds $\max\{\max_{1 \leqslant t \leqslant T} \mathcal{S}_t, \max_{k \neq j} \max_{1 \leqslant t \leqslant T} \mathcal{S}_t^{(k)}\}$. Since all such probabilities sum to one, we conclude that $\mathbb{P}(\max_{1 \leqslant t \leqslant T} \mathcal{S}_t > \mathfrak{z}) = 1/(J+1)$, provided that there are no change points. We took $J = 9$ in the experiments with changes in mean and in variance. In the third experiment, where distribution change of observations transforms but the first two moments remain unchanged, we took $J = 4$. The thresholds of other algorithms were tuned in a similar fashion. In other words, the running length of all the algorithms was at least 150 with probability 0.9 in the experiments on the first and the second data sets and with probability 0.8 in the experiments on the third data set.

The setup was as follows. In each example, we sampled an artificial sequence 10 times and computed the detection delays for Algorithm 1 with different classes $\mathcal{F}$ and for the competitors (CUSUM, KLIEP and kernel change point detection with M-statistic) for each realization. Table 1 displays the average detection delay for each method. In all the synthetic experiments, the weights of the neural net-

[1] https://github.com/npuchkin/contrastive_change_point_detection

work were optimized via the PyTorch implementation of the Adam method (Kingma and Ba, 2015) with 50 epochs and the learning rate 0.1. During the first 20 iterations, we collected the observations for further training, and the test statistic was not computed. We also slightly adjusted the test statistic $\mathcal{S}_t$: instead of maximizing $\mathcal{T}_{\tau,t}(\widehat{f}_{\tau,t})$ over the whole set $\{1,\dots,t-1\}$, we took the maximum with respect to $\tau \in \{10,11,\dots,t-10\}$. This simple trick helped us to reduce the detection delay. The hyperparameters of KLIEP and M-statistic-based kernel change point methods were tuned in a way to minimize the average detection delay while keeping the running length at least 150 with high probability. The information about the thresholds and hyperparameters is collected in Table 4 (see Section E in the appendix).

Example 1: mean shift detection in a Gaussian sequence model. We generated a univariate Gaussian sequence of length 150. The first 75 observations had the Gaussian distribution $\mathcal{N}(0,\sigma^2)$ with $\sigma = 0.1$ and the other 75 were i.i.d. $\mathcal{N}(\mu,\sigma^2)$ with $\mu = 0.2$ and the same σ. Besides the class of neural networks, we considered the class of polynomials of degree $p = 1$ and the linear span of $\{1, \sin(2\pi x)\}$ (that is, the linear span of the first $q = 2$ elements of the Fourier basis).

Example 2: variance change detection in a Gaussian sequence model. In the second example, we sampled 75 independent Gaussian random variables $\mathcal{N}(0,\sigma_0^2)$ with $\sigma_0 = 0.1$ and 75 random variables with the distribution $\mathcal{N}(0,\sigma^2)$, $\sigma = 0.3$, so the expectation of all the random variables was the same. CUSUM is not applicable in this case. The parameters p and q were set to 2 and 3, respectively.

Example 3: distributional change. Finally, we checked the ability of our procedure to adapt to distributional changes. For this purpose, we generated a sequence of 150 independent random variables where the first 75 had the uniform distribution on $[-\sigma/\sqrt{3}, \sigma/\sqrt{3}]$ with $\sigma = 0.1$ and the other 75 were drawn from the Gaussian distribution $\mathcal{N}(0,\sigma^2)$. The parameters of the uniform distribution were chosen in a way to match the first two moments of the Gaussian distribution. In this case, CUSUM is not applicable. The parameters p and q were set to 5 and 6, respectively.

According to Table 1, Algorithm 1 with $\mathcal{F}$ equal to polynomials is the most efficient method to detect a change point amongst competitors. KLIEP takes the third place in the first experiment and the second place in two others. CUSUM detects the shift in mean extremely fast but it is not applicable to more comlicated scenarios. We would like to note that Algorithm 1 with $\mathcal{F}$ equal to neural networks performs almost as good as KLIEP.

Table 1: Detection delays of Algorithm 1 (with three variants of the class $\mathcal{F}$: polynomials (poly), linear span of the Fourier basis (Fourier) and neural networks (NN)), KLIEP, kernel change point with M-statistic, and CUSUM on synthetic data sets. Two best results are boldfaced.

METHOD	EX. 1	EX. 2	EX. 3
Alg. 1			
+ poly	$\mathbf{6.7 \pm 2.0}$	$\mathbf{16.4 \pm 8.1}$	$\mathbf{41.2 \pm 28.3}$
Alg. 1			
+ Fourier	7.6 ± 2.1	44.1 ± 16.8	62.0 ± 22.1
Alg. 1			
+ NN	9.4 ± 1.6	19.8 ± 8.4	59.1 ± 19.0
KLIEP	9.0 ± 3.5	$\mathbf{19.6 \pm 18.9}$	$\mathbf{43.0 \pm 32.1}$
M-statistic	10.4 ± 3.4	51.1 ± 27.3	65.3 ± 18.1
CUSUM	$\mathbf{5.0 \pm 2.0}$	–	–

4.2 Speech records analysis

We used CENSREC-1-C[2] data in the Speech Resource Consortium (SRC) corpora provided by National Institute of Informatics (NII) to test the algorithm in practical tasks. The data set contains a clean speech record (MAH_clean) and the same record corrupted with noise of different magnitude (MAH_N1_SNR20, MAH_N1_SNR15). We preprocessed the data as follows. First, we normalized the data. Next, the audio track was split into 5 segments with a single change from silence/noise to speech, and then each 10-th observation was taken. The true change point values were set on the MAH_clean data set and used in the noisy versions of the record. Examples of behaviour of test statistics for different methods are exposed in Figure 4 (see Section E in the appendix).

Table 2: Detection delays of Algorithm 1 (with three variants of the class $\mathcal{F}$), KLIEP, kernel change point detector with M-statistic, and CUSUM on the CENSREC-1-C speech records (the clean one and two corrupted with noise with SNR 20 and SNR 15). Two best results are boldfaced.

METHOD	CLEAN	SNR 20	SNR 15
Algorithm 1			
+ polynomials	$\mathbf{3.2 \pm 4.0}$	$\mathbf{3.8 \pm 2.6}$	$\mathbf{6.5 \pm 7.8}$
Algorithm 1			
+ Fourier basis	8.2 ± 6.2	8.2 ± 2.2	11.0 ± 6.9
Algorithm 1			
+ neural networks	7.5 ± 6.8	$\mathbf{3.0 \pm 1.2}$	9.0 ± 5.9
KLIEP	7.8 ± 9.7	17.0 ± 13.7	10.2 ± 8.8
M-statistic	$\mathbf{3.2 \pm 3.1}$	10.5 ± 7.9	$\mathbf{4.2 \pm 4.3}$

As in the experiments with the artificial data sets, we considered Algorithm 1 with three variants of the class $\mathcal{F}$:

[2] http://research.nii.ac.jp/src/en/CENSREC-1-C.html

polynomials of degree 9, the linear span of the first 10 elements of the Fourier basis, and the class of fully connected feed-forward neural networks with architecture $(1, 2, 3, 1)$ and ReLU activations. We used the Adam optimizer with 200 epochs and the learning rate 0.1 to tune the parameters of the neural network. We computed detection delays for each algorithm on each of 5 segments. The results are reported in Table 2, the corresponding thresholds and the values of hyperparameters are shown in Table 5 (see Section E in the appendix).

Similarly to the experiments with artificial data, Algorithm 1 with polynomial $\mathcal{F}$ proved its efficiency being amongst two best methods in all the cases. The main difference with the artificial experiments is that kernel change point detection method with M-statistic behaves much better, while KLIEP shows poor performance, compared to Algorithm 1 with polynomials and neural networks and the M-statistic-based method.

4.3 Activity change recognition

In this section, we apply Algorithm 1 to detect changes in a user's physical activity. In our experiments, we took a part of the data set WISDM (Weiss et al., 2019), containing 3-dimensional measurements of a smartphone accelerometer, measured at a rate 20Hz. We preprocessed the data set, taking only each 20-th observation. Nevertheless, even after such a reduction the length of the time series was over 3000. The observations are displayed in Figure 2. During the measurement period, the user changed a kind of activity 17 times, i.e. the time series contained 17 change points. Our goal was to detect them as soon as possible.

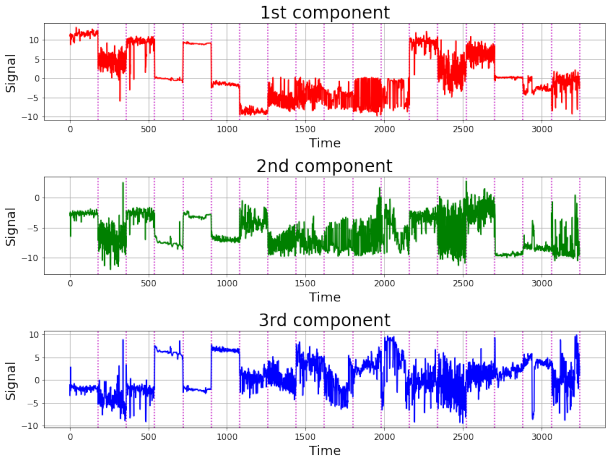


Figure 2: Three-dimensional time series from the WISDM data set.

We applied Algorithm 1 with two variants of the class $\mathcal{F}$: a linear class and a class of three-layered fully-connected feed-forward neural networks with an architecture $(1, 2, 3, 1)$. As before, we compared our procedure with KLIEP and the kernel change point detector with M-statistic. After the parameter tuning, we chose the bandwidth in KLIEP and M-statistic equal to 0.5. We took the thresholds to be equal to $\mathfrak{z} = 22$, $\mathfrak{z} = 20$, $\mathfrak{z} = 75$, and $\mathfrak{z} = 7.5$ in Algorithm 1 with a linear class $\mathcal{F}$, Algorithm 1 with neural networks, the kernel change point detector, and KLIEP, respectively. After that, we computed the number of false alarms and the average detection delay. The results are presented in Table 3. The plots of the test statistics are shown in Figure 5. We would like to note that the kernel change point detector missed the second change point. The average detection delay for M-statistic is computed based on all the change points, except for the second one. According to Table 3, Algorithm 1 makes less false alarms while having a smaller average detection delay.

Table 3: The number of false alarms (FA) and the average detection delays (DD) of Algorithm 1 (with two variants of the class $\mathcal{F}$), KLIEP, and the kernel change point detector with M-statistic on the WISDM data set. Best results are boldfaced.

METHOD	FA	DD
Algorithm 1 + linear class	**4**	23.1 ± 12.3
Algorithm 1 + neural networks	**4**	$\mathbf{16.4 \pm 5.7}$
KLIEP	5	35.2 ± 49.6
M-statistic	7	30.2 ± 44.4

5 CONCLUSION AND FUTURE DIRECTIONS

We suggested a novel online change point detection procedure which is suitable for both parametric and nonparametric scenarios. We derived high probability bounds on the running length and the detection delay of the algorithm. As a consequence, we obtained the first non-asymptotic bound for online change point detection via neural networks. We also conducted numerical experiments on artificial and real-world data illustrating efficiency of the proposed method.

Further research in this direction may include consideration of nonstationary post-change observations as in (Liang et al., 2021). Besides, one may try to improve the dependence on B in the upper bound (7) using improper estimators instead of $\widehat{f}$. In (Foster et al., 2018), the authors showed that a proper regularization leads to a doubly-exponential improvement in the dependence on B in the problem of logistic regression.

Note on societal impacts The paper is mostly of theoretical nature. The presented results and the change point detection algorithm itself have no negative societal impact.

Acknowledgements

The publication was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University №70-2021-00139. Nikita Puchkin is a Young Russian Mathematics award winner and would like to thank its sponsors and jury. The authors are grateful to the anonymous referees for careful reading of the paper and for their valuable and constructive remarks that improved the quality of this work.

References

S. Arlot, A. Celisse, and Z. Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, 20(162):1–56, 2019.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

D. Belomestny, L. Iosipoi, Q. Paris, and N. Zhivotovskiy. Empirical variance minimization with applications in variance reduction and optimal control. *Bernoulli*, 28 (2):1382–1407, 2022.

G. Biau, K. Bleakley, and D. M. Mason. Long signal change-point detection. *Electronic Journal of Statistics*, 10(2):2097–2123, 2016.

Y. Cao, L. Xie, Y. Xie, and H. Xu. Sequential change-point detection via online convex optimization. *Entropy*, 20 (2):108, 2018.

W.-C. Chang, C.-L. Li, Y. Yang, and B. Póczos. Kernel change-point detection with auxiliary deep generative models. In *International Conference on Learning Representations*, 2019.

Y. Chen, T. Wang, and R. J. Samworth. High-dimensional, multiscale online changepoint detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84:234–266, 2022.

L. Chu and H. Chen. Sequential change-point detection for high-dimensional and non-euclidean data. *IEEE Transactions on Signal Processing*, 70:4498–4511, 2022.

R. Corradin, L. Danese, and A. Ongaro. Bayesian nonparametric change point detection for multivariate time series with missing observations. *International Journal of Approximate Reasoning*, 143:26–43, 2022.

R. C. Dalang and A. N. Shiryaev. A quickest detection problem with an observation cost. *The Annals of Applied Probability*, 25(3):1475–1512, 2015.

H. Dehling, K. Vuk, and M. Wendler. Change-point detection based on weighted two-sample U-statistics. *Electronic Journal of Statistics*, 16(1):862–891, 2022.

H. Dette and J. Gösmann. A likelihood ratio approach to sequential change point detection for a general class of parameters. *Journal of the American Statistical Association*, 115(531):1361–1377, 2020.

L. Dümbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 29(1): 124–152, 2001.

B. Eichinger and C. Kirch. A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018.

F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079, 2019.

D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 167–208, 2018.

D. Garreau and S. Arlot. Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12(2): 4440–4486, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Grover, J. Song, A. Kapoor, K. Tran, A. Agarwal, E. J. Horvitz, and S. Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

Z. Harchaoui, E. Moulines, and F. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

A. Hero. Geometric entropy minimization (gem) for anomaly detection and localization. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

M. Hushchyn and A. Ustyuzhanin. Generalization of change-point detection in time series data based on direct density ratio estimation. *J. Comput. Sci.*, 53:Paper No. 101385, 8, 2021.

M. Hushchyn, K. Arzymatov, and D. Derkach. Online neural networks for change-point detection. Preprint, arXiv:2010.01388, 2020.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, 2015.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

K. K. Korkas and P. Fryzlewicz. Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, 27(1):287–311, 2017.

M. N. Kurt, Y. Yilmaz, and X. Wang. Real-time nonparametric anomaly detection in high-dimensional settings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:2463–2479, 2021.

M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes*. Berlin: Springer, reprint of the 1991 edition, 2011.

S. Li, Y. Xie, H. Dai, and L. Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Y. Liang, A. G. Tartakovsky, and V. V. Veeravalli. Quickest change detection with non-stationary post-change observations. Preprint, arXiv:2110.01581, 2021.

S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

M. Londschien, P. Bühlmann, and S. Kovàcs. Random forests for change point detection. Preprint, arXiv:2205.04997, 2022.

O. H. Madrid Padilla, Y. Yu, D. Wang, and A. Rinaldo. Optimal nonparametric change point analysis. *Electronic Journal of Statistics*, 15(1):1154–1201, 2021.

D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 06 1954.

E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3-4):523–527, 12 1955.

F. Pein, H. Sieling, and A. Munk. Heterogeneous change point inference. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 79(4):1207–1227, 2017.

M. Pollak and A. G. Tartakovsky. Optimality properties of the Shiryaev-Roberts procedure. *Statistica Sinica*, 19(4): 1729–1739, 2009.

A. Rinaldo, D. Wang, Q. Wen, R. Willett, and Y. Yu. Localizing changes in high-dimensional regression models. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2089–2097, 2021.

S. W. Roberts. A comparison of some control chart procedures. *Technometrics*, 8(3):411–430, 1966.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

J. Shin, A. Ramdas, and A. Rinaldo. E-detectors: a nonparametric framework for online changepoint detection. Preprint, arXiv:2203.03532, 2022.

A. N. Shiryaev. The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Mathematics. Doklady*, 2:795–799, 1961.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8: 22–46, 1963.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Y.-W. Sun, K. Papagiannouli, and V. Spokoiny. High dimensional change-point detection: a complete graph approach. Preprint, arXiv:2203.08709, 2022.

A. G. Tartakovsky, M. Pollak, and A. S. Polunchenko. Third-order asymptotic optimality of the generalized shiryaev-roberts changepoint detection procedures. *Theory of Probability & Its Applications*, 56(3):457–484, 2012.

M. K. Titsias, J. Sygnowski, and Y. Chen. Sequential changepoint detection in neural networks with checkpoints. *Statistics and Computing*, 32(2):26, 2022.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

D. Wang, Y. Yu, and A. Rinaldo. Univariate mean change point detection: penalization, CUSUM and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961, 2020.

G. M. Weiss, K. Yoneda, and T. Hayajneh. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access*, 7:133190–133202, 2019.

M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25 (5):1324–1370, 2013.

Y. Yu, S. Chatterjee, and H. Xu. Localising change points in piecewise polynomials of general degrees. Preprint, arXiv:2007.09910, 2020a.

Y. Yu, O. H. M. Padilla, D. Wang, and A. Rinaldo. A note on online change point detection. Preprint, arXiv:2006.03283, 2020b.

C. Zou, G. Yin, L. Feng, and Z. Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002, 2014.

A SOME PROPERTIES OF SUB-GAUSSIAN RANDOM VARIABLES

In this section, we provide some useful properties of sub-Gaussian random variables. For any random variable ξ, its Orlicz norm ψ_2 is defined as

$$\|\xi\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}e^{\xi^2/t^2} \leqslant 2\right\}.$$

Random variables with a finite ψ_2-norm are usually called sub-Gaussian, because the tails of their distributions decay as $O\left(e^{-t^2/\|\xi\|_{\psi_2}^2}\right)$. Indeed, by the definition of the Orlicz norm, we have

$$\mathbb{P}\left(|\xi| > t\right) \leqslant \frac{\mathbb{E}e^{|\xi|/\|\xi\|_{\psi_2}^2}}{e^{t^2/\|\xi\|_{\psi_2}^2}} \leqslant 2e^{-\frac{t^2}{\|\xi\|_{\psi_2}^2}}, \quad \text{for all } t > 0. \tag{11}$$

In its turn, the inequality (11) yields an upper bound on the L_p-norm of ξ. For any $p \geqslant 1$, it holds that

$$\begin{aligned}\mathbb{E}|\xi|^p &= \int\limits_0^{+\infty} \mathbb{P}\left(|\xi|^p \geqslant u\right) \mathrm{d}u \leqslant 2\int\limits_0^{+\infty} e^{-\frac{u^{2/p}}{\|\xi\|_{\psi_2}^2}} \mathrm{d}u \\ &= p\|\xi\|_{\psi_2}^p \int\limits_0^{+\infty} v^{p/2-1}e^{-v}\mathrm{d}v = p\|\xi\|_{\psi_2}^p \Gamma\left(\frac{p}{2}\right) = 2\|\xi\|_{\psi_2}^p \Gamma\left(\frac{p}{2}+1\right),\end{aligned} \tag{12}$$

where $\Gamma(\cdot)$ is the gamma function. There are several equivalent definitions of sub-Gaussian variables. A reader can find them, for instance, in (Vershynin, 2018, Proposition 2.5.2). In our proofs, we deal with sums of independent sub-Gaussian random variables and use the following property.

Proposition A.1 (Vershynin (2018), Proposition 2.6.1)**.** *Let $\xi_1, \ldots, \xi_n$ be independent centered sub-Gaussian random variables. Then their sum $S_n = \xi_1 + \ldots + \xi_n$ is also sub-Gaussian, and its Orlicz norm satisfies the inequality*

$$\|S_n\|_{\psi_2}^2 \lesssim \sum_{i=1}^n \|\xi_i\|_{\psi_2}^2.$$

It is also worth mentioning that, according to (Vershynin, 2018, Lemma 2.6.8), the centering does not increase the ψ_2-norm too much. That is, for any sub-Gaussian random variable ξ, it holds that

$$\|\xi - \mathbb{E}\xi\|_{\psi_2} \lesssim \|\xi\|_{\psi_2}. \tag{13}$$

B SOME PROPERTIES OF NEURAL NETWORKS

This section collects some useful properties of feed-forward neural networks with ReLU activations. Let us recall that $\mathsf{NN}(L, \mathcal{A}, s)$ denotes the class of neural networks with L hidden layers, architecture $\mathcal{A}$, and at most s non-zero weights. The next theorem from Schmidt-Hieber (2020) concerns approximation properties of neural networks.

Theorem B.1 (Schmidt-Hieber (2020), Theorem 5)**.** *For any $f^* \in \mathcal{H}^\beta([0,1]^p, H)$ and any $N, m \in \mathbb{N}$, there exists a neural network $f \in \mathsf{NN}(L, \mathcal{A}, s)$ with*

$$L = 8 + (m+5)\left(1 + \lceil \log_2(p \vee \beta)\rceil\right)$$

hidden layers, the architecture

$$\mathcal{A} = (p, 6(\lceil\beta\rceil + p)N, \ldots, 6(\lceil\beta\rceil + p)N, 1),$$

and the number of non-zero parameters

$$s \leqslant 141(p + \beta + 1)^{3+p} N(m+6),$$

such that

$$\|f - f^*\|_{L_\infty([0,1]^p)} \leqslant (1 + p^2 + \beta^2)6^p(2H+1)N2^{-m} + 3^\beta H N^{-\beta/p}.$$

Taking a sufficiently deep and wide enough neural network, one can approximate any function from $\mathcal{H}^\beta([0,1]^p, H)$ with the desired accuracy. On the other hand, Schmidt-Hieber established the following upper bound on the covering number of $\mathsf{NN}(L, \mathcal{A}, s)$.

Lemma B.2 (Schmidt-Hieber (2020), Lemma 5)**.** *For any $L \in \mathbb{Z}_+, \mathcal{A} \in \mathbb{N}^{L+2}, s \in \mathbb{N}$, the covering number of the class $\mathsf{NN}(L, \mathcal{A}, s)$ with respect to the $L_\infty([0,1]^p)$-norm satisfies the inequality*

$$\log \mathcal{N}(\mathsf{NN}(L, \mathcal{A}, s), L_\infty([0,1]^p), \varepsilon) \leqslant (s+1) \log \left(\frac{2(L+1)V^2}{\varepsilon} \right), \quad \textit{for all } \varepsilon > 0,$$

where $V = \prod_{j=0}^{L+1}(1 + a_j)$.

C PROOFS OF THE MAIN RESULTS

C.1 Proof of Lemma 2.1

Let

$$D^*(x) = \frac{e^{f^*(x)}}{1 + e^{f^*(x)}} = \frac{\mathsf{p}(x)}{\mathsf{p}(x) + \mathsf{q}(x)}.$$

With the introduced notation, it holds that

$$f^*(x) - \ln\left(\frac{1 + e^{f^*(x)}}{2}\right) = \ln(2D^*(x)) \quad \text{and} \quad -\ln\left(\frac{1 + e^{f^*(x)}}{2}\right) = \ln(2 - 2D^*(x)).$$

If $\tau^* = \tau$, then we obtain that

$$\begin{aligned}
\mathbb{E}\mathcal{T}_{\tau,t}(f^*) &= \frac{\tau(t-\tau)}{t}\left[\int \ln(2D^*(x))\mathsf{p}(x)\mathrm{d}\mathfrak{m} + \int \ln(2 - 2D^*(x))\mathsf{q}(x)\mathrm{d}\mathfrak{m}\right] \\
&= \frac{\tau(t-\tau)}{t}\left[\int \ln\left(\frac{2\mathsf{p}(x)}{\mathsf{p}(x) + \mathsf{q}(x)}\right)\mathsf{p}(x)\mathrm{d}\mathfrak{m} + \int \ln\left(\frac{2\mathsf{q}(x)}{\mathsf{p}(x) + \mathsf{q}(x)}\right)\mathsf{q}(x)\mathrm{d}\mathfrak{m}\right] \\
&= \frac{2\tau(t-\tau)}{t}\left[\mathrm{KL}\left(\mathsf{p}, \frac{\mathsf{p}+\mathsf{q}}{2}\right) + \mathrm{KL}\left(\mathsf{q}, \frac{\mathsf{p}+\mathsf{q}}{2}\right)\right] \\
&\equiv \frac{2\tau(t-\tau)\,\mathrm{JS}(\mathsf{p}, \mathsf{q})}{t}.
\end{aligned}$$

Fix a function f, introduce $D(x) = e^{f(x)}/(1 + e^{f(x)})$ and note that

$$\begin{aligned}
\mathbb{E}\mathcal{T}_{\tau,t}(f^*) - \mathbb{E}\mathcal{T}_{\tau,t}(f) &= \frac{\tau(t-\tau)}{t}\left[\int \ln\left(\frac{D^*(x)}{D(x)}\right)\mathsf{p}(x)\mathrm{d}\mathfrak{m} + \int \ln\left(\frac{1 - D^*(x)}{1 - D(x)}\right)\mathsf{q}(x)\mathrm{d}\mathfrak{m}\right] \\
&= \frac{\tau(t-\tau)}{t}\left[\int D^*(x)\ln\left(\frac{D^*(x)}{D(x)}\right)(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right. \\
&\qquad \left. + \int (1 - D^*(x))\ln\left(\frac{1 - D^*(x)}{1 - D(x)}\right)(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right].
\end{aligned}$$

Substituting $D^*(x)$ and $D(x)$ by $e^{f^*(x)}/(1 + e^{f^*(x)})$ and $1/(1 + e^{f^*(x)})$, respectively, we obtain that

$$\begin{aligned}
\mathbb{E}\mathcal{T}_{\tau,t}(f^*) - \mathbb{E}\mathcal{T}_{\tau,t}(f) &= \frac{\tau(t-\tau)}{t}\left[\int \frac{e^{f^*(x)}}{1 + e^{f^*(x)}}(f^*(x) - f(x))(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right] \\
&\quad - \frac{\tau(t-\tau)}{t}\left[\int \frac{e^{f^*(x)}}{1 + e^{f^*(x)}}\ln\left(\frac{1 + e^{f^*(x)}}{1 + e^{f(x)}}\right)(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right] \\
&\quad - \frac{\tau(t-\tau)}{t}\left[\int \frac{1}{1 + e^{f^*(x)}}\ln\left(\frac{1 + e^{f^*(x)}}{1 + e^{f(x)}}\right)(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right] \qquad (14) \\
&= \frac{\tau(t-\tau)}{t}\left[\int \frac{e^{f^*(x)}}{1 + e^{f^*(x)}}(f^*(x) - f(x))(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right] \\
&\quad - \frac{\tau(t-\tau)}{t}\left[\int \ln\left(\frac{1 + e^{f^*(x)}}{1 + e^{f(x)}}\right)(\mathsf{p}(x) + \mathsf{q}(x))\mathrm{d}\mathfrak{m}\right].
\end{aligned}$$

Consider a function $g : \mathbb{R}^2 \to \mathbb{R}$, defined as

$$g(u, v) = \frac{(u - v)e^u}{1 + e^u} - \ln\left(\frac{1 + e^u}{1 + e^v}\right).$$

Note that, for any $u, v \in \mathbb{R}$, we have $g(u, u) = 0$,

$$\frac{\partial g(u, v)}{\partial v}\bigg|_{v=u} = \left[-\frac{e^u}{1 + e^u} + \frac{e^v}{1 + e^v}\right]\bigg|_{v=u} = 0, \quad \text{and} \quad \frac{\partial^2 g(u, v)}{\partial v^2} = \frac{e^v}{(1 + e^v)^2} \leqslant \frac{1}{4}.$$

Hence, for any $u, v \in \mathbb{R}$, it holds that

$$g(u, v) \leqslant \frac{(u - v)^2}{8}.$$

Applying this inequality to the right-hand side of (14), we obtain that

$$\begin{aligned}\mathbb{E}\mathcal{T}_{\tau,t}(f^*) - \mathbb{E}\mathcal{T}_{\tau,t}(f) &\leqslant \frac{\tau(t - \tau)}{8t}\left[\int (f^*(x) - f(x))^2 \left(\mathsf{p}(x) + \mathsf{q}(x)\right)\mathrm{d}\mathfrak{m}\right] \\ &\leqslant \frac{\tau(t - \tau)}{8t}\left(\|f^* - f\|^2_{L_2(\mathsf{p})} + \|f^* - f\|^2_{L_2(\mathsf{q})}\right).\end{aligned}$$

Taking into account that $\mathbb{E}\mathcal{T}_{\tau,t}(f^*) = 2\tau(t - \tau)\,\mathrm{JS}(\mathsf{p}, \mathsf{q})/t$, we finally get

$$\mathbb{E}\mathcal{T}_{\tau,t}(f) \geqslant \frac{2\tau(t - \tau)}{t}\left(\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \frac{1}{16}\|f - f^*\|^2_{L_2(\mathsf{p})} - \frac{1}{16}\|f - f^*\|^2_{L_2(\mathsf{q})}\right).$$

C.2 Proof of Theorem 2.3

Lemma C.1. *Let a function f take its values in $[-B, B]$. Assume that $X_1, \ldots, X_t$ are independent and identically distributed. Then, for any $\tau \in \{1, \ldots, t - 1\}$, it holds that*

$$\frac{\tau(t - \tau)\mathbb{E}f^2(X_1)}{t} \leqslant \frac{-\mathbb{E}\mathcal{T}_{\tau,t}(f)}{\varkappa}.$$

where $\mathcal{T}_{\tau,t}(f)$ is defined in (4) *and*

$$\varkappa = \min\left\{\frac{e^B}{(1 + e^B)^2}, \frac{e^{-B}}{(1 + e^{-B})^2}\right\}.$$

Moreover, we have $\mathrm{Var}(\mathcal{T}_{\tau,t}(f)) \leqslant \tau(t - \tau)\mathbb{E}f^2(X_1)/t \leqslant -\mathbb{E}\mathcal{T}_{\tau,t}(f)/\varkappa$.

In the proof of Theorem 2.3, we use an approach based on local Rademacher complexities (see, for instance, Bartlett et al. (2005)). The inequality $\mathrm{Var}(\mathcal{T}_{\tau,t}(f)) \leqslant -\mathbb{E}\mathcal{T}_{\tau,t}(f)/\varkappa$ will allow us to get the so-called fast rates of convergence. However, note that

$$\frac{t - \tau}{t}\sum_{s=1}^{\tau}\left[f(X_s) - \ln\left(\frac{1 + e^{f(X_s)}}{2}\right)\right] \quad \text{nor} \quad \frac{\tau}{t}\sum_{s=\tau+1}^{t}\ln\left(\frac{1 + e^{f(X_s)}}{2}\right)$$

do not have the properties of $\mathcal{T}_{\tau,t}(f)$, discussed in Lemma C.1. This means that, in order to exploit the curvature of $\mathbb{E}\mathcal{T}_{\tau,t}(f)$ with respect to f, we must study both terms in (4) simultaneously. The problem is that the terms in the right-hand side of (4) are not identically distributed (though independent). We must slightly modify the argument of Bartlett et al. (2005) to overcome this issue.

Introduce a parameter $r > 0$ to be specified later. For any $f \in \mathcal{F}$, define

$$\mathsf{k}(f) = \min\left\{m \in \mathbb{Z}_+ : 4^m r \geqslant \frac{\tau(t - \tau)}{t}\mathbb{E}f^2(X_1)\right\}$$

and consider the empirical process $4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)$, $f \in \mathcal{F}$. The next lemma allows us to bound the expectation of

$$\sup_{f \in \mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right].$$

Lemma C.2. *Grant Assumption 2.2. Then there exists an absolute constant $C > 0$ such that*

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right]\\ &\leqslant C\left(\sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)} + Bd\log\left(\frac{A\tau(t-\tau)}{rt}\right)\right)\end{aligned} \tag{15}$$

The proof of Lemma C.2 is based on the bracketing entropy chaining argument (Han et al., 2019, Lemma 7). Denote the right-hand side of (15) by $\Phi(r)$. Talagrand's concentration inequality (Klein and Rio, 2005, Theorem 1.1), combined with the result of Lemma C.2, yields that

$$\sup_{f\in\mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right] \leqslant 2\Phi(r) + \sqrt{2r\log(1/\delta)} + 4B\log(1/\delta)$$

on an event E_1, such that $\mathbb{P}(E_1) \geqslant 1-\delta$. Here we used the fact that, for any $f \in \mathcal{F}$, it holds that

$$\mathrm{Var}\left(4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right) = \frac{1}{16^{\mathsf{k}(f)}}\mathrm{Var}\left(\mathcal{T}_{\tau,t}(f)\right) \leqslant \frac{\tau(t-\tau)}{4^{\mathsf{k}(f)}t}\mathbb{E}f^2(X_1) \leqslant r.$$

Take

$$r = \max\left\{\frac{128Cd}{\varkappa}\left(\frac{4Cd}{\varkappa}\vee B\right)\log\left(\frac{A\tau(t-\tau)\varkappa}{16Ctd(B\wedge 4Cd/\varkappa)}\right), \left(\frac{8}{\varkappa}\vee B\right)\frac{16\log(1/\delta)}{\varkappa}\right\} \tag{16}$$

and consider two cases. For all functions $f \in \mathcal{F}$, satisfying the inequality $\tau(t-\tau)\mathbb{E}f^2(X_1) \leqslant rt$, we have $\mathsf{k}(f) = 0$ and then, on the event E_1, it holds that

$$\begin{aligned}\mathcal{T}_{\tau,t}(f) &\leqslant \mathbb{E}\mathcal{T}_{\tau,t}(f) + 2\Phi(r) + \sqrt{2r\log(1/\delta)} + 4B\log(1/\delta)\\ &\leqslant 2\Phi(r) + \sqrt{2r\log(1/\delta)} + 4B\log(1/\delta)\\ &\lesssim \frac{d}{\varkappa}\log\left(\frac{A\varkappa^2\tau(t-\tau)}{td}\right) + Bd\log\left(\frac{A\varkappa\tau(t-\tau)}{tBd}\right) + \left(\frac{1}{\varkappa}\vee B\right)\log(1/\delta).\end{aligned}$$

Here we used the fact that $\mathbb{E}\mathcal{T}_{\tau,t}(f) \leqslant 0$ for all $f \in \mathcal{F}$ (follows from Lemma C.1). Otherwise, due to the definition of $\mathsf{k}(f)$, it holds that

$$\frac{-\mathbb{E}\mathcal{T}_{\tau,t}(f)}{\varkappa} \geqslant \frac{\tau(t-\tau)\mathbb{E}f^2(X_1)}{t} \geqslant 4^{\mathsf{k}(f)-1}r$$

and, hence,

$$\mathcal{T}_{\tau,t}(f) \leqslant 4^{\mathsf{k}(f)}\left(-\frac{\varkappa r}{4} + 2\Phi(r) + \sqrt{2r\log(1/\delta)} + 4B\log(1/\delta)\right)$$

on E_1. For r given by (16), we have

$$-\frac{\varkappa r}{16} \geqslant \max\left\{\Phi(r), \sqrt{2r\log(1/\delta)}, 4B\log(1/\delta)\right\}.$$

Thus, we obtain that $\mathcal{T}_{\tau,t}(f) \leqslant 0$ on E_1 for all $f \in \mathcal{F}$ such that $\mathsf{k}(f) \geqslant 1$. Hence, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\mathcal{T}_{\tau,t}(f) \lesssim \frac{d}{\varkappa}\log\left(\frac{A\varkappa^2\tau(t-\tau)}{td}\right) + Bd\log\left(\frac{A\varkappa\tau(t-\tau)}{tBd}\right) + \left(\frac{1}{\varkappa}\vee B\right)\log(1/\delta).$$

The expression in the right-hand side can be simplified if one takes into account that $\varkappa \geqslant 0.5e^{-B}$:

$$\sup_{f\in\mathcal{F}}\mathcal{T}_{\tau,t}(f) \lesssim de^B\left[B + \log\left(\frac{A\tau(t-\tau)}{td}\right)\right] + e^B\log(1/\delta).$$

C.3 Proof of Theorem 2.6

As in the proof of Theorem 2.3, we use the peeling and reweighting argument. Introduce a parameter $r > 0$ to be specified later. Recall that, for any $f \in \mathcal{F}$,

$$\mathsf{k}(f) = \min\left\{m \in \mathbb{Z}_+ : 4^m r \geqslant \frac{\tau(t-\tau)}{t}\mathbb{E}f^2(X_1)\right\},$$

and, for any $b \geqslant a > 0$, define

$$\mathcal{F}(a,b) = \left\{f \in \mathcal{F} : \frac{at}{\tau(t-\tau)} \leqslant \mathbb{E}f^2(X_1) \leqslant \frac{bt}{\tau(t-\tau)}\right\}.$$

Then it holds that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right] \\
&\leqslant \max\left\{\sup_{f\in\mathcal{F}(0,r)}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right],\right. \\
&\qquad\left.\max_{j\in\mathbb{Z}_+}\sup_{f\in\mathcal{F}(4^{j-1}r,4^j r)}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f) - \mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right]\right\} \\
&\leqslant \max_{j\in\mathbb{Z}_+}\left\{4^{-j+1}\sup_{f\in\mathcal{F}(0,4^j r)}\left[\mathcal{T}_{\tau,t}(f) - \mathbb{E}\mathcal{T}_{\tau,t}(f)\right]\right\}.
\end{aligned}$$

Fix any $f, g \in \mathcal{F}$. Due to the centering lemma (Vershynin, 2018, Lemma 2.6.8) (see also the inequality (13)), it holds that

$$\left\|\mathcal{T}_{\tau,t}(f) - \mathbb{E}\mathcal{T}_{\tau,t}(f) - \mathcal{T}_{\tau,t}(g) + \mathbb{E}\mathcal{T}_{\tau,t}(g)\right\|_{\psi_2} \lesssim \left\|\mathcal{T}_{\tau,t}(f) - \mathcal{T}_{\tau,t}(g)\right\|_{\psi_2}.$$

since the probability measure is clear from context, we write ψ_2, instead of $\psi_2(\mathsf{p})$, in this proof to avoid the abuse of notation. For any $f \in \mathcal{F}$, let us represent

$$\mathcal{T}_{\tau,t}(f) = \frac{\tau(t-\tau)}{t}\mathcal{P}_{\tau,t}(f) + \frac{\tau(t-\tau)}{t}\mathcal{Q}_{\tau,t}(f),$$

where

$$\mathcal{P}_{\tau,t}(f) = \frac{1}{\tau}\sum_{s=1}^{\tau}\left[f(X_s) - \ln\left(\frac{1+e^{f(X_s)}}{2}\right)\right]$$

and

$$\mathcal{Q}_{\tau,t}(f) = -\frac{1}{t-\tau}\sum_{s=\tau+1}^{t}\ln\left(\frac{1+e^{f(X_s)}}{2}\right).$$

The triangle inequality yields that

$$\left\|\mathcal{T}_{\tau,t}(f) - \mathcal{T}_{\tau,t}(g)\right\|_{\psi_2} \leqslant \frac{\tau(t-\tau)}{t}\left\|\mathcal{P}_{\tau,t}(f) - \mathcal{P}_{\tau,t}(g)\right\|_{\psi_2} + \frac{\tau(t-\tau)}{t}\left\|\mathcal{Q}_{\tau,t}(f) - \mathcal{Q}_{\tau,t}(g)\right\|_{\psi_2}.$$

Applying Proposition A.1 to $\mathcal{P}_{\tau,t}(f) - \mathcal{P}_{\tau,t}(g)$ and $\mathcal{Q}_{\tau,t}(f) - \mathcal{Q}_{\tau,t}(g)$, we obtain that

$$\left\|\mathcal{P}_{\tau,t}(f) - \mathcal{P}_{\tau,t}(g)\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{\tau}}\left\|\ln\left(\frac{e^{f(X_1)}}{1+e^{f(X_1)}}\right) - \ln\left(\frac{e^{g(X_1)}}{1+e^{g(X_1)}}\right)\right\|_{\psi_2}$$

and

$$\left\|\mathcal{Q}_{\tau,t}(f) - \mathcal{Q}_{\tau,t}(g)\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{t-\tau}}\left\|\ln\left(1+e^{f(X_1)}\right) - \ln\left(1+e^{g(X_1)}\right)\right\|_{\psi_2}.$$

Moreover, since the maps $y \mapsto (y - \log(1+e^y))$ and $y \mapsto \log(1+e^y)$ are 1-Lipschitz, we have

$$\left\|\ln\left(\frac{e^{f(X_1)}}{1+e^{f(X_1)}}\right) - \ln\left(\frac{e^{g(X_1)}}{1+e^{g(X_1)}}\right)\right\|_{\psi_2} \leqslant \|f-g\|_{\psi_2}$$

and, similarly,

$$\left\|\ln\left(1+e^{f(X_1)}\right)-\ln\left(1+e^{g(X_1)}\right)\right\|_{\psi_2}\leqslant\|f-g\|_{\psi_2}.$$

Hence,

$$\begin{aligned}\|\mathcal{T}_{\tau,t}(f)-\mathcal{T}_{\tau,t}(g)\|_{\psi_2}&\lesssim\left(\frac{(t-\tau)\sqrt{\tau}}{t}+\frac{\tau\sqrt{t-\tau}}{t}\right)\|f-g\|_{\psi_2}\\&\lesssim\sqrt{\frac{t}{\tau(t-\tau)}}\,\|f-g\|_{\psi_2}\\&\leqslant L\sqrt{\frac{t}{\tau(t-\tau)}}\,\|f-g\|_{L_2(\mathsf{p})}\,,\end{aligned}$$

and we can apply a corollary of (Ledoux and Talagrand, 2011, Theorem 11.2 and eq. (11.3)) (see the discussion in (Ledoux and Talagrand, 2011, Theorem p. 302)): for any $j\in\mathbb{Z}_+$, it holds that

$$\begin{aligned}&\left\|\sup_{f\in\mathcal{F}(0,4^jr)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)]\right\|_{\psi_2}\\&\lesssim L\sqrt{\frac{t}{\tau(t-\tau)}}\int\limits_0^{\mathcal{D}(\mathcal{F}(0,4^jr),L_2(\mathsf{p}))}\sqrt{\log\mathcal{N}(\mathcal{F}(0,4^jr),L_2(\mathsf{p}),u)}\mathrm{d}u\\&\leqslant L\sqrt{\frac{t}{\tau(t-\tau)}}\int\limits_0^{4^jrt/\tau/(t-\tau)}\sqrt{d\log\left(\frac{A}{\varepsilon}\right)}\mathrm{d}u\\&\lesssim L\sqrt{4^jrd\log\left(\frac{A\tau(t-\tau)}{4^jrt}\right)}.\end{aligned}$$

This and (11) imply that, for any $\delta\in(0,1)$ and for any $j\in\mathbb{Z}_+$, with probability at least $1-2^{-j-1}\delta$, it holds that

$$\sup_{f\in\mathcal{F}(0,4^jr)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)]\lesssim L\sqrt{4^jrd\log\left(\frac{A\tau(t-\tau)}{4^jrt}\right)\log\left(\frac{2^{j+2}}{\delta}\right)}.$$

Applying the union bound, we obtain that, for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)-\mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right]&\leqslant\max_{j\in\mathbb{Z}_+}\left\{4^{-j+1}\sup_{f\in\mathcal{F}(0,4^jr)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)]\right\}\\&\lesssim\max_{j\in\mathbb{Z}_+}\left\{2^{-j}L\sqrt{rd\log\left(\frac{A\tau(t-\tau)}{4^jrt}\right)\log\left(\frac{2^{j+2}}{\delta}\right)}\right\}\\&\lesssim L\sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)\log(1/\delta)}.\end{aligned}\tag{17}$$

Let C be a hidden constant in (17) and take

$$r=\frac{16C^2L^2d}{\kappa^2}\log\left(\frac{A\kappa^2\tau(t-\tau)}{16C^2L^2td}\right)\log(1/\delta).$$

From now on, we restrict our attention on an event E_2, where (17) holds. As in the proof of Theorem 2.3, consider two cases. First, if a function $f\in\mathcal{F}$ satisfies $\tau(t-\tau)\mathbb{E}f^2(X_1)\leqslant rt$, then $\mathsf{k}(f)=0$ and, hence,

$$\begin{aligned}\mathcal{T}_{\tau,t}(f)\leqslant\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)&\lesssim L\sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)\log(1/\delta)}\\&\lesssim\frac{L^2d}{\kappa}\log\left(\frac{A\kappa^2\tau(t-\tau)}{L^2td}\right)\log(1/\delta).\end{aligned}$$

On the other hand, the inequality $\mathsf{k}(f) \geqslant 1$ means that $\tau(t-\tau)\mathbb{E}f^2(X_1) \geqslant 4^{\mathsf{k}(f)-1}rt$. The next lemma relates the expectations of $f^2(X_1)$ and $\mathcal{T}_{\tau,t}(f)$.

Lemma C.3. *Assume that a function class $\mathcal{F}$ is L-sub-Gaussian with respect to $X_1 \sim \mathsf{p}$. Fix any $t \in \mathbb{N}$, $\tau \in \{1, \ldots, t-1\}$ and let $X_2, \ldots, X_t$ be i.i.d. copies of X_1. Then, for any $f \in \mathcal{F}$, it holds that*

$$\frac{\tau(t-\tau)}{t}\mathbb{E}f^2(X_1) \leqslant \frac{-\mathbb{E}\mathcal{T}_{\tau,t}}{\kappa}, \quad \textit{where} \quad \kappa = \frac{1}{2}\exp\left\{-\mathcal{D}(\mathcal{F}, \psi_2)\sqrt{2\ln(4L\sqrt{2})}\right\}.$$

Lemma C.3 immediately implies that

$$\frac{4^{-\mathsf{k}(f)}\mathbb{E}\mathcal{T}_{\tau,t}}{\kappa} \leqslant -r/4,$$

and then

$$\mathcal{T}_{\tau,t}(f) \leqslant -\frac{\kappa r}{4} + CL\sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)\log(1/\delta)} \leqslant 0.$$

Thus, on the event E_2, it holds that

$$\begin{aligned}\sup_{f\in\mathcal{F}}\mathcal{T}_{\tau,t}(f) &\lesssim \frac{L^2 d}{\kappa}\log\left(\frac{A\kappa^2\tau(t-\tau)}{L^2 td}\right)\log(1/\delta)\\ &\lesssim L^2 d e^{\mathcal{D}(\mathcal{F},\psi_2)\sqrt{2\log(4L\sqrt{2})}}\left[\mathcal{D}(\mathcal{F},\psi_2)\sqrt{\log L} + \log\left(\frac{A\tau(t-\tau)}{L^2 td}\right)\right]\log(1/\delta).\end{aligned}$$

C.4 Proof of Theorem 2.7

Theorem 2.3 and the union bound yield that, in the stationary regime, with probability at least $1-\delta$

$$\max_{1\leqslant t\leqslant T}\mathcal{S}_t \leqslant Cde^B\left[B + \log\left(\frac{AT}{d}\right)\right] + Ce^B\log(T/\delta),$$

where C is an absolute positive constant. Hence, if $X_1, \ldots, X_T$ are i.i.d. random elements and

$$\mathfrak{z} = Cde^B\left[B + \log\left(\frac{AT}{d}\right)\right] + Ce^B\log(T/\delta),$$

then Algorithm 1 does not stop on the first T iterations with probability at least $1-\delta$.

On the other hand, let $f^\circ \in \operatorname{argmin}_{f\in\mathcal{F}}\|f - \log(\mathsf{p}/\mathsf{q})\|_{L_2(\mathsf{p}+\mathsf{q})}$. Bernstein's inequality implies that, for any fixed $t \in \mathbb{N}$, with probability at least $1-\delta$,

$$\begin{aligned}\mathcal{T}_{\tau^*,t}(f^\circ) &> \mathbb{E}\mathcal{T}_{\tau^*,t}(f^\circ) - \sqrt{2\mathrm{Var}(\mathcal{T}_{\tau^*,t}(f^\circ))\log(1/\delta)} - 3B\log(1/\delta)\\ &> \frac{2\tau^*(t-\tau^*)}{t}\left(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \frac{\rho^2(\mathcal{F})}{16}\right) - B\left(\sqrt{\frac{2(t-\tau^*)\tau^*\log(1/\delta)}{t}} + 3\log(1/\delta)\right).\end{aligned}$$

Here we used the fact that $\log((1+e^u)/2) \leqslant |u|$ for all $u \in \mathbb{R}$, which yields

$$\begin{aligned}\mathrm{Var}(\mathcal{T}_{\tau^*,t}(f^\circ)) &\leqslant \frac{(t-\tau^*)^2\tau^*}{t}\mathbb{E}\log^2\left(\frac{2e^{f^\circ(X_1)}}{1+e^{f^\circ(X_1)}}\right) + \frac{(t-\tau^*)\tau^{*2}}{t}\mathbb{E}\log^2\left(\frac{2}{1+e^{f^\circ(X_t)}}\right)\\ &\leqslant \frac{(t-\tau^*)^2\tau^*}{t}\mathbb{E}\left(f^\circ(X_1)\right)^2 + \frac{(t-\tau^*)\tau^{*2}}{t}\mathbb{E}\left(f^\circ(X_1)\right)^2\\ &\leqslant \frac{B^2(t-\tau^*)\tau^*}{t}.\end{aligned}$$

Let t° be the smallest positive integer, satisfying the inequality

$$\frac{\tau^*(t-\tau^*)}{t} \geqslant \frac{\tau_{\min}}{2},$$

where

$$\frac{\tau_{\min}}{2} = \left\lceil \frac{B^2 \log(1/\delta)}{2(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{3B\log(1/\delta) + \mathfrak{z}}{2(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)} \right\rceil + 1.$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
\mathcal{S}_{t^\circ} &\geqslant \mathcal{T}_{\tau,t^\circ}(f^\circ) \\
&> \frac{2\tau(t-\tau)}{t}\left(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \frac{\rho^2(\mathcal{F})}{16}\right) - B\left(\sqrt{\frac{2(t-\tau)\tau\log(1/\delta)}{t}} + 3\log(1/\delta)\right) \\
&\geqslant \mathfrak{z}.
\end{aligned}$$

Thus, on this event, the stopping time of Algorithm 1 does not exceed t°. This implies that

$$\frac{\tau^*(\widehat{t}-\tau^*)}{\widehat{t}} \leqslant \frac{\tau^*(t^\circ+1-\tau^*)}{t^\circ+1} < \frac{\tau_{\min}}{2}.$$

Note that, due to the conditions of Theorem 2.7, it holds that $\tau^* \geqslant \tau_{\min}$. Hence, with probability at least $1-\delta$,

$$\widehat{t} - \tau^* \leqslant \frac{\tau_{\min}\tau^*}{2(\tau^* - \tau_{\min}/2)} \leqslant \tau_{\min} \lesssim \frac{B^2\log(1/\delta)}{(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{B\log(1/\delta)+\mathfrak{z}}{\mathrm{JS}(\mathsf{p},\mathsf{q}) - \rho^2(\mathcal{F})/16}.$$

C.5 Proof of Theorem 2.9

The proof of Theorem 2.9 is similar to the one of Theorem 2.7 but relies on 2.6, rather than on 2.3. Theorem 2.6 and the union bound imply that, in the stationary regime, the exists such $C > 0$ that, with probability at least $1-\delta$,

$$\max_{1\leqslant t\leqslant T} \mathcal{S}_t \leqslant CL^2 d e^{\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{2\log(4L\sqrt{2})}}\left[\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{\log L} + \log\left(\frac{A\tau(t-\tau)}{L^2 t d}\right)\right]\log(T/\delta).$$

Hence, if

$$\mathfrak{z} = CL^2 d e^{\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{2\log(4L\sqrt{2})}}\left[\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p}))\sqrt{\log L} + \log\left(\frac{A\tau(t-\tau)}{L^2 t d}\right)\right]\log(T/\delta),$$

then the running length of Algorithm 1 exceeds T with probability at least $1-\delta$.

On the other hand, due to (11), for $f^\circ \in \operatorname{argmin}_{f\in\mathcal{F}} \|f - \log(\mathsf{p}/\mathsf{q})\|_{L_2(\mathsf{p}+\mathsf{q})}$ and any $t \in \mathbb{N}$, with probability at least $1-\delta$, it holds that

$$\mathcal{T}_{\tau^*,t}(f^\circ) - \mathbb{E}\mathcal{T}_{\tau^*,t}(f^\circ) > -\|\mathcal{T}_{\tau^*,t}(f^\circ) - \mathbb{E}\mathcal{T}_{\tau^*,t}(f^\circ)\|_{\psi_2}\sqrt{\log\frac{2}{\delta}}.$$

According to Proposition A.1 and (13),

$$\begin{aligned}
&\|\mathcal{T}_{\tau^*,t}(f^\circ) - \mathbb{E}\mathcal{T}_{\tau^*,t}(f^\circ)\|_{\psi_2} \lesssim \|\mathcal{T}_{\tau^*,t}(f^\circ)\|_{\psi_2} \\
&\lesssim \frac{(t-\tau^*)\tau^*}{t}\left(\frac{1}{\sqrt{\tau^*}}\left\|\ln\left(\frac{2e^{f^\circ}}{1+e^{f^\circ}}\right)\right\|_{\psi_2(\mathsf{p})} + \frac{1}{\sqrt{t-\tau^*}}\left\|\ln\left(\frac{1+e^{f^\circ}}{2}\right)\right\|_{\psi_2(\mathsf{q})}\right) \\
&\lesssim \sqrt{\frac{(t-\tau^*)\tau^*}{t}}\Big(\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F},\psi_2(\mathsf{q}))\Big).
\end{aligned}$$

Hence, there exists an absolute constant $c > 0$ such that, for any $t \in \mathbb{N}$, with probability at least $1-\delta$,

$$\begin{aligned}
\mathcal{S}_t &\geqslant \mathcal{T}_{\tau^*,t}(f^\circ) > \mathbb{E}\mathcal{T}_{\tau^*,t}(f^\circ) - c\sqrt{\frac{(t-\tau^*)\tau^*\log(1/\delta)}{t}}\Big(\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F},\psi_2(\mathsf{q}))\Big) \\
&\geqslant \frac{2\tau^*(t-\tau^*)}{t}\left(\mathrm{JS}(\mathsf{p},\mathsf{q}) - \frac{\rho^2(\mathcal{F})}{16}\right) - c\sqrt{\frac{(t-\tau^*)\tau^*\log(1/\delta)}{t}}\Big(\mathcal{D}(\mathcal{F},\psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F},\psi_2(\mathsf{q}))\Big)
\end{aligned}$$

Let t° be the smallest positive integer, satisfying the inequality

$$\frac{\tau^*(t-\tau^*)}{t} \geqslant \frac{\tau_{\min}}{2},$$

where

$$\frac{\tau_{\min}}{2} = \left\lceil \frac{c^2 \left[\mathcal{D}(\mathcal{F}, \psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F}, \psi_2(\mathsf{q}))\right]^2 \log(1/\delta)}{(\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{\mathfrak{z}}{\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \rho^2(\mathcal{F})/16} \right\rceil + 1.$$

Then, with probability at least $1 - \delta$, we have

$$\begin{aligned}
\mathcal{S}_{t^\circ} &\geqslant \mathcal{T}_{\tau, t^\circ}(f^\circ) \\
&> \frac{2\tau(t-\tau)}{t}\left(\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \frac{\rho^2(\mathcal{F})}{16}\right) - c\sqrt{\frac{(t-\tau^*)\tau^*}{t}\log(1/\delta)}\Big(\mathcal{D}(\mathcal{F}, \psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F}, \psi_2(\mathsf{q}))\Big) \\
&\geqslant \mathfrak{z}.
\end{aligned}$$

Thus, on this event, the stopping time of Algorithm 1 does not exceed t°. This implies that

$$\frac{\tau^*(\widehat{t} - \tau^*)}{\widehat{t}} \leqslant \frac{\tau^*(t^\circ + 1 - \tau^*)}{t^\circ + 1} < \frac{\tau_{\min}}{2}.$$

Note that, due to the conditions of Theorem 2.9, it holds that $\tau^* \geqslant \tau_{\min}$. Hence, with probability at least $1 - \delta$,

$$\begin{aligned}
\widehat{t} - \tau^* &\leqslant \frac{\tau_{\min}\tau^*}{2(\tau^* - \tau_{\min}/2)} \leqslant \tau_{\min} \\
&\lesssim \frac{\left[\mathcal{D}(\mathcal{F}, \psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F}, \psi_2(\mathsf{q}))\right]^2 \log(1/\delta)}{(\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \rho^2(\mathcal{F})/16)^2} + \frac{\mathfrak{z}}{\mathrm{JS}(\mathsf{p}, \mathsf{q}) - \rho^2(\mathcal{F})/16}.
\end{aligned}$$

C.6 Proof of Corollary 3.1

Take the smallest positive integers m, N, satisfying the inequalities

$$3^\beta H N^{-\beta/p} \leqslant \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}/2 \quad \text{and} \quad (1 + p^2 + \beta^2)6^p(2H+1)N2^{-m} \leqslant \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}/2,$$

and consider the class $\mathsf{NN}(L, \mathcal{A}, s)$ with

$$L = 8 + (m+5)\left(1 + \lceil \log_2(p \vee \beta) \rceil\right) \tag{18}$$

hidden layers, the architecture

$$\mathcal{A} = (p, 6(\lceil\beta\rceil + p)N, \ldots, 6(\lceil\beta\rceil + p)N, 1), \tag{19}$$

and the number of non-zero parameters

$$s = 141(p + \beta + 1)^{3+p} N(m+6). \tag{20}$$

According to Theorem B.1, there exists $f \in \mathsf{NN}(L, \mathcal{A}, s)$ such that

$$\|f - \ln(\mathsf{p}/\mathsf{q})\|_{L_\infty([0,1]^p)} \leqslant \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}.$$

Note that, since $\log(\mathsf{p}/\mathsf{q}) \in \mathcal{H}^\beta([0,1]^p, H)$ the L_∞-norm of such f does not exceed $H + \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}$. Thus, $f \in \mathsf{NN}_B(L, \mathcal{A}, s)$ for $L, \mathcal{A}$, and s, given by (18), (19), (5), respectively, and for any $B > H + \sqrt{\mathrm{JS}(\mathsf{p}, \mathsf{q})}$.

On the other hand, due to Lemma B.2, for any $\varepsilon > 0$, the covering number of $\mathsf{NN}_B(L, \mathcal{A}, s)$ with respect to the L_∞-norm fulfils

$$\log \mathcal{N}(\mathsf{NN}_B(L, \mathcal{A}, s), L_\infty([0,1]^p), \varepsilon) \leqslant (s+1)\log\left(\frac{2(L+1)p(6\lceil\beta\rceil + 6p)^L N^L}{\varepsilon}\right).$$

Hence, $\mathsf{NN}_B(L, \mathcal{A}, s)$ satisfies (2.2) with $d = s + 1$ and $A = 2(L+1)p(6\lceil\beta\rceil + 6p)^L N^L$. Taking into account that

$$L \lesssim \log(1/\,\mathrm{JS}(\mathsf{p}, \mathsf{q})), \quad N \lesssim \mathrm{JS}(\mathsf{p}, \mathsf{q})^{-p/(2\beta)}, \quad s \lesssim \mathrm{JS}(\mathsf{p}, \mathsf{q})^{-p/(2\beta)} \log(1/\,\mathrm{JS}(\mathsf{p}, \mathsf{q})),$$

and substituting these bounds into Theorem 2.7, we obtain that if one chooses $\mathfrak{z}$ according to

$$\mathfrak{z} = \frac{Ce^B \log(1/\,\mathrm{JS}(\mathsf{p}, \mathsf{q}))\left[B + \log(1/\,\mathrm{JS}(\mathsf{p}, \mathsf{q}))\log T\right]}{\mathrm{JS}(\mathsf{p}, \mathsf{q})^{p/(2\beta)}} + Ce^B \log(T/\delta)$$

with a proper constant $C > 0$ and runs Algorithm 1 is run with $\mathcal{F} = \mathsf{NN}_B(L, \mathcal{A}, s)$, where $L, \mathcal{A}$, and s are defined in (18), (19), and (5), respectively, then, with probability at least $1 - \delta$, its running length in the stationary regime is at least T. Otherwise, if $\tau^* < \infty$, then, with probability at least $1 - \delta$, the stopping time $\widehat{t}$ of Algorithm 1 satisfies

$$\widehat{t} - \tau^* \lesssim \frac{e^B \log(1/\operatorname{JS}(\mathsf{p}, \mathsf{q}) \left[B + \log(1/\operatorname{JS}(\mathsf{p}, \mathsf{q})) \log T\right]}{\operatorname{JS}(\mathsf{p}, \mathsf{q})^{\frac{2\beta+p}{2\beta}}} + e^B \log(T/\delta) + \frac{B^2 \log(1/\delta)}{\operatorname{JS}(\mathsf{p}, \mathsf{q})^2}.$$

This finishes the proof of Corollary 3.1.

C.7 Proof of Corollary 3.2

First, show that $\mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{q})) \lesssim \|\Sigma^{-1/2}\mu\|$. Let X_1 be a Gaussian random vector with zero mean and the covariance Σ. Then, for any $w \in \mathbb{R}^p$, such that $\|\Sigma^{1/2} w\| \leqslant \|\Sigma^{-1/2}\mu\|$, we have $w^\top X_1 \sim \mathcal{N}(0, w^\top \Sigma w)$ and

$$\|w^\top X_1\|_{\psi_2(\mathsf{p})} \lesssim \sqrt{w^\top \Sigma w} \leqslant \|\Sigma^{-1/2}\mu\|.$$

At the same time, for any $b \in \mathbb{R}$, such that $|b| \leqslant \mu^\top \Sigma^{-1} \mu$, it holds that

$$\|b\|_{\psi_2(\mathsf{p})} \lesssim \mu^\top \Sigma^{-1}\mu \lesssim \|\Sigma^{-1/2}\mu\|,$$

where the last inequality is due to the fact $\|\Sigma^{-1/2}\mu\| \leqslant \ln(4/3)$. Hence, by the triangle inequality,

$$\|w^\top X_1 + b\|_{\psi_2(\mathsf{p})} \leqslant \|w^\top X_1\|_{\psi_2(\mathsf{p})} + \|b\|_{\psi_2(\mathsf{p})} \lesssim \|\Sigma^{-1/2}\mu\|$$

for all $w \in \mathbb{R}^p, b \in \mathbb{R}$, such that $\|\Sigma^{1/2} w\| \leqslant \|\Sigma^{-1/2}\mu\|$, $|b| \leqslant \mu^\top \Sigma^{-1}\mu$. Thus, $\mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{p})) \lesssim \|\Sigma^{-1/2}\mu\|$. Similarly, $\mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{q})) \lesssim \|\Sigma^{-1/2}\mu\|$.

Second, show that $\mathcal{F}_{\text{lin}}$ satisfies Assumption 2.5. For any $w_1, w_2 \in \mathbb{R}^p$ and $b_1, b_2 \in \mathbb{R}$, it holds that

$$\|w_1^\top X_1 + b_1 - w_2^\top X_1 - b_2\|^2_{L_2(\mathsf{p})} = (w_1 - w_2)^\top \Sigma (w_1 - w_2) + \|b_1 - b_2\|^2.$$

This yields that, if $\mathcal{W}$ is an ε-net of the ellipsoid $\{w : \|\Sigma^{1/2} w\| \leqslant \|\Sigma^{-1/2}\mu\|\}$ and $\mathcal{B}$ is an ε-net of the segment $[-\mu^\top \Sigma^{-1}\mu, \mu^\top \Sigma^{-1}\mu]$, then the set

$$\left\{ f_{w,b}(x) = w^\top x + b : w \in \mathcal{W}, b \in \mathcal{B} \right\}$$

is an $(\varepsilon\sqrt{2})$-net of $\mathcal{F}_{\text{lin}}$. Thus, we conclude that $\log \mathcal{N}(\mathcal{F}_{\text{lin}}, L_2(\mathsf{p}), \varepsilon) \lesssim p \log(\mu^\top \Sigma^{-1}\mu/\varepsilon)$ for any $\varepsilon > 0$.

It only remains to show that $\operatorname{JS}(\mathsf{p}, \mathsf{q}) \gtrsim \mu^\top \Sigma^{-1}\mu$. Then, substituting the obtained bounds on $\mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{p})) \vee \mathcal{D}(\mathcal{F}_{\text{lin}}, \psi_2(\mathsf{q}))$, $\mathcal{N}(\mathcal{F}_{\text{lin}}, L_2(\mathsf{p}), \varepsilon)$, and $\operatorname{JS}(\mathsf{p}, \mathsf{q})$ into the statement of Theorem 2.9, we get the assertion of Corollary 3.2.

The rest of this section is devoted to the proof of the inequality $\operatorname{JS}(\mathsf{p}, \mathsf{q}) \gtrsim \mu^\top \Sigma^{-1}\mu$. By the definition of $\operatorname{JS}(\mathsf{p}, \mathsf{q})$,

$$\operatorname{JS}(\mathsf{p}, \mathsf{q}) = \frac{\operatorname{KL}(\mathsf{p}, (\mathsf{p}+\mathsf{q})/2) + \operatorname{KL}(\mathsf{q}, (\mathsf{p}+\mathsf{q})/2)}{2}.$$

Consider the first term:

$$\operatorname{KL}\left(\mathsf{p}, \frac{\mathsf{p}+\mathsf{q}}{2}\right) = \mathbb{E}_{\xi \sim \mathsf{p}} \log \frac{2\mathsf{p}}{\mathsf{p}+\mathsf{q}} = -\mathbb{E}_{\xi \sim \mathsf{p}} \log \frac{1 + e^{\mu^\top \Sigma^{-1}\xi - \mu^\top \Sigma^{-1}\mu/2}}{2}$$

Let us introduce $\eta = \mu^\top \Sigma^{-1}\xi - \mu^\top \Sigma^{-1}\mu/2 \sim \mathcal{N}(-\mu^\top \Sigma^{-1}\mu/2, \mu^\top \Sigma^{-1}\mu)$. Since the second derivative of the map $u \mapsto \log((1+e^u)/2)$ does not exceed $1/4$, Jensen's inequality implies that

$$\begin{aligned}
\operatorname{KL}\left(\mathsf{p}, \frac{\mathsf{p}+\mathsf{q}}{2}\right) &= -\mathbb{E} \ln \frac{1+e^\eta}{2} \\
&\geqslant -\ln \frac{1+e^{\mathbb{E}\eta}}{2} - \frac{\operatorname{Var}(\eta)}{8} \\
&= -\ln \frac{1+e^{-\mu^\top \Sigma^{-1}\mu/2}}{2} - \frac{\mu^\top \Sigma^{-1}\mu}{8}.
\end{aligned}$$

Consider the function $g(u) = -\ln((1+e^{-u})/2)$. Note that $g(0) = 0$, $g(1) > 1/3$, and g is concave on $[0,1]$. This yields that $g(u) \geqslant u/3$ for all $u \in [0,1]$. Hence,

$$\mathrm{KL}\left(\mathsf{p}, \frac{\mathsf{p}+\mathsf{q}}{2}\right) \geqslant -\ln\frac{1+e^{-\mu^\top\Sigma^{-1}\mu/2}}{2} - \frac{\mu^\top\Sigma^{-1}\mu}{8} \geqslant \frac{\mu^\top\Sigma^{-1}\mu}{6} - \frac{\mu^\top\Sigma^{-1}\mu}{8} = \frac{\mu^\top\Sigma^{-1}\mu}{24}.$$

Similarly, one can prove that

$$\mathrm{KL}\left(\mathsf{q}, \frac{\mathsf{p}+\mathsf{q}}{2}\right) \gtrsim \mu^\top\Sigma^{-1}\mu,$$

and, therefore, $\mathrm{JS}(\mathsf{p},\mathsf{q}) \gtrsim \mu^\top\Sigma^{-1}\mu$.

D PROOFS OF AUXILIARY RESULTS

D.1 Proof of Lemma C.1

It holds that

$$-\mathbb{E}\mathcal{T}_{\tau,t}(f) = \frac{\tau(t-\tau)}{t}\mathbb{E}\left[-f(X_1) + 2\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right].$$

Consider a function $G : [-B, B] \to \mathbb{R}$, defined as

$$G(u) = -u + 2\ln\left(\frac{1+e^u}{2}\right).$$

Direct calculations show that $G(0) = G'(0) = 0$ and

$$G''(u) = \frac{2e^u}{(1+e^u)^2} \geqslant 2\varkappa, \quad \text{for all } u \in [-B, B],$$

where

$$\varkappa = \min\left\{\frac{e^B}{(1+e^B)^2}, \frac{e^{-B}}{(1+e^{-B})^2}\right\}.$$

Hence, using Taylor's expansion, we obtain that $G(u) \geqslant \varkappa u^2$ for all $u \in [-B, B]$. This yields that

$$\frac{\varkappa\tau(t-\tau)\mathbb{E}f^2(X_1)}{t} \leqslant -\mathbb{E}\mathcal{T}_{\tau,t}(f). \tag{21}$$

To prove the second part of the lemma, note that

$$\begin{aligned}\mathrm{Var}(\mathcal{T}_{\tau,t}) &= \frac{\tau(t-\tau)^2}{t^2}\mathrm{Var}\left[f(X_1) - \ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right] + \frac{\tau^2(t-\tau)}{t^2}\mathrm{Var}\left[\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right] \\ &\leqslant \frac{\tau(t-\tau)^2}{t^2}\mathbb{E}\left[f(X_1) - \ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right]^2 + \frac{\tau^2(t-\tau)}{t^2}\mathbb{E}\left[\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right]^2.\end{aligned}$$

Since the functions $G_1(u) = u - \ln[(1+e^u)/2]$ and $G_2(u) = \ln[(1+e^u)/2]$ are 1-Lipschitz and $G_1(0) = G_2(0) = 0$, we have

$$\mathrm{Var}(\mathcal{T}_{\tau,t}) \leqslant \frac{\tau(t-\tau)^2}{t^2}\mathbb{E}f^2(X_1) + \frac{\tau^2(t-\tau)}{t^2}\mathbb{E}f^2(X_1) = \frac{\tau(t-\tau)}{t}\mathbb{E}f^2(X_1) \leqslant \frac{-\mathbb{E}\mathcal{T}_{\tau,t}(f)}{\varkappa},$$

where the last inequality is due to (21).

D.2 Proof of Lemma C.2

Let us recall that, for any $b \geqslant a > 0$, $\mathcal{F}(a,b)$ is defined as

$$\mathcal{F}(a,b) = \left\{f \in \mathcal{F} : \frac{at}{\tau(t-\tau)} \leqslant \mathbb{E}f^2(X_1) \leqslant \frac{bt}{\tau(t-\tau)}\right\}.$$

Then it holds that

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)-\mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right]\\
&\leqslant \mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)-\mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right]\\
&\quad+\sum_{j=0}^{\infty}\mathbb{E}\sup_{f\in\mathcal{F}(4^j r,4^{j+1}r)}\left[4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)-\mathbb{E}4^{-\mathsf{k}(f)}\mathcal{T}_{\tau,t}(f)\right] \qquad (22)\\
&\leqslant \mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)\right]+\sum_{j=0}^{\infty}4^{-j}\,\mathbb{E}\sup_{f\in\mathcal{F}(0,4^{j+1}r)}\left[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)\right].
\end{aligned}$$

For any $f\in\mathcal{F}$, let us represent $\mathcal{T}_{\tau,t}(f)$ as a sum of two terms:

$$\mathcal{T}_{\tau,t}(f)=\frac{\tau(t-\tau)}{t}\mathcal{P}_{\tau,t}(f)+\frac{\tau(t-\tau)}{t}\mathcal{Q}_{\tau,t}(f),$$

where

$$\mathcal{P}_{\tau,t}(f)=\frac{1}{\tau}\sum_{s=1}^{\tau}\left[f(X_s)-\ln\left(\frac{1+e^{f(X_s)}}{2}\right)\right]$$

and

$$\mathcal{Q}_{\tau,t}(f)=-\frac{1}{t-\tau}\sum_{s=\tau+1}^{t}\ln\left(\frac{1+e^{f(X_s)}}{2}\right).$$

Then, for any $r>0$, it holds that

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)\right]&\leqslant\frac{\tau(t-\tau)}{t}\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{P}_{\tau,t}(f)-\mathbb{E}\mathcal{P}_{\tau,t}(f)\right]\\
&\quad+\frac{\tau(t-\tau)}{t}\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{Q}_{\tau,t}(f)-\mathbb{E}\mathcal{Q}_{\tau,t}(f)\right].
\end{aligned}$$

We apply the following lemma to bound the expectations of the suprema in the right-hand side.

Lemma D.1 (Han et al. (2019), Lemma 7; in this form, Belomestny et al. (2022), Lemma A.6)**.** *Let $\xi_1,\dots,\xi_n$, $n\in\mathbb{N}$, be independent copies of a random variable $\xi\sim\mathsf{P}$, and let $\mathcal{H}$ be a class of functions taking its values in $[-B,B]$. Suppose that, for all $0<u\leqslant B$,*

$$\mathcal{N}_{[\,]}(\mathcal{H},L_2(\mathsf{P}),u)\leqslant\left(\frac{A}{\varepsilon}\right)^d$$

for some positive constants A and d. Then

$$\mathbb{E}\sup_{h\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n}h(\xi_i)-\mathbb{E}h(\xi)\right]\lesssim\sqrt{\frac{d\sigma^2}{n}\log\left(\frac{A}{\sigma}\right)}+\frac{Bd}{n}\log\left(\frac{A}{\sigma}\right),$$

where $\sigma^2=\sup\limits_{h\in\mathcal{H}}\mathbb{E}h^2(\xi)$.

Note that the maps $y\mapsto(y-\log(1+e^y))$ and $y\mapsto\log(1+e^y)$ are monotonously increasing and 1-Lipschitz. This yields that if f belongs to a bracket $[f_1,f_2]$ of size ε, then $(f-\log(1+e^f))$ is in the bracket $[f_1-\log(1+e^{f_1}),f_2-\log(1+e^{f_2})]$ of size at most ε and, similarly, $\log(1+e^f)$ belongs to the bracket $[\log(1+e^{f_1}),\log(1+e^{f_2})]$ of size at most ε. In other words, a monotonous 1-Lipschitz map does not change the bracketing number. Thus, it holds that

$$\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{P}_{\tau,t}(f)-\mathbb{E}\mathcal{P}_{\tau,t}(f)\right]\lesssim\sqrt{\frac{rtd}{\tau^2(t-\tau)}\log\left(\frac{A\tau(t-\tau)}{rt}\right)}+\frac{Bd}{\tau}\log\left(\frac{A\tau(t-\tau)}{rt}\right)$$

and

$$\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}\left[\mathcal{Q}_{\tau,t}(f)-\mathbb{E}\mathcal{Q}_{\tau,t}(f)\right]\lesssim\sqrt{\frac{rtd}{\tau(t-\tau)^2}\log\left(\frac{A\tau(t-\tau)}{rt}\right)}+\frac{Bd}{t-\tau}\log\left(\frac{A\tau(t-\tau)}{rt}\right).$$

Therefore, due to the definitions of $\mathcal{T}_{\tau,t}(f)$, $\mathcal{P}_{\tau,t}(f)$, and $\mathcal{Q}_{\tau,t}(f)$,

$$\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)] \lesssim \sqrt{\frac{rd(t-\tau)}{t}\log\left(\frac{A\tau(t-\tau)}{rt}\right)}+\sqrt{\frac{rd\tau}{t}\log\left(\frac{A\tau(t-\tau)}{rt}\right)} + Bd\log\left(\frac{A\tau(t-\tau)}{rt}\right).$$

Using the inequality $\sqrt{a}+\sqrt{b}\leqslant\sqrt{2(a+b)}$, which holds for all non-negative a and b, we obtain that

$$\mathbb{E}\sup_{f\in\mathcal{F}(0,r)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)] \lesssim \sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)}+Bd\log\left(\frac{A\tau(t-\tau)}{rt}\right). \tag{23}$$

Similarly, we can prove that

$$\mathbb{E}\sup_{f\in\mathcal{F}(0,4^{j+1}r)}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)] \lesssim \sqrt{4^{j+1}rd\log\left(\frac{A\tau(t-\tau)}{4^{j+1}rt}\right)} + Bd\log\left(\frac{A\tau(t-\tau)}{4^{j+1}rt}\right). \tag{24}$$

Substituting the bounds (23) and (24) into the inequality (22), we get that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}[\mathcal{T}_{\tau,t}(f)-\mathbb{E}\mathcal{T}_{\tau,t}(f)] &\lesssim \sum_{j=0}^{\infty}4^{-j}\left[\sqrt{4^jrd\log\left(\frac{A\tau(t-\tau)}{4^jrt}\right)}+Bd\log\left(\frac{A\tau(t-\tau)}{4^jrt}\right)\right]\\ &\lesssim \sqrt{rd\log\left(\frac{A\tau(t-\tau)}{rt}\right)}+Bd\log\left(\frac{A\tau(t-\tau)}{rt}\right).\end{aligned}$$

D.3 Proof of Lemma C.3

Similarly to Lemma C.1, we have

$$-\mathbb{E}\mathcal{T}_{\tau,t}(f)=\frac{\tau(t-\tau)}{t}\mathbb{E}\left[-f(X_1)+2\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right].$$

Consider a function $G:\mathbb{R}\to\mathbb{R}$, defined as

$$G(u)=-u+2\ln\left(\frac{1+e^u}{2}\right).$$

Direct calculations show that $G(0)=G'(0)=0$ and

$$G''(u)=\frac{2e^u}{(1+e^u)^2}\geqslant\frac{e^u}{2(1\vee e^u)^2}=\frac{e^u\wedge e^{-u}}{2}=\frac{e^{-|u|}}{2}.$$

Using Taylor's expansion with an integral remainder, we obtain that

$$G(u)\geqslant u^2\int\limits_0^1 G''(yu)(1-y)\mathrm{d}y\geqslant\frac{u^2}{2}\int\limits_0^1 e^{-y|u|}(1-y)\mathrm{d}y\geqslant\frac{u^2e^{-|u|}}{4}.$$

This yields

$$\mathbb{E}\left[-f(X_1)+2\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right]\geqslant\frac{\mathbb{E}f^2(X_1)e^{-|f(X_1)|}}{4}.$$

Denote $\xi=|f(X_1)|$ and note that $\|\xi\|_{\psi_2}=\|f(X_1)\|_{\psi_2}$. Then it holds that

$$\mathbb{E}\left[\xi^2e^{-\xi}\right]\geqslant\mathbb{E}\left[\xi^2e^{-\xi}\mathbb{I}(\xi\leqslant a)\right]\geqslant e^{-a}\mathbb{E}\left[\xi^2\mathbb{I}(\xi\leqslant a)\right]=e^{-a}\mathbb{E}\xi^2-e^{-a}\mathbb{E}\left[\xi^2\mathbb{I}(\xi>a)\right].$$

Consider the second term in the right-hand side. Due to the Cauchy-Schwarz inequality, it holds that

$$\mathbb{E}\left[\xi^2\mathbb{I}(\xi > a)\right] \leqslant \sqrt{\mathbb{E}\xi^4\mathbb{P}(\xi > a)}.$$

Applying the inequality (12) for L_p-norms of sub-Gaussian random variables, we obtain that $\mathbb{E}\xi^4 \leqslant 4\|\xi\|_{\psi_2}^4$ and, thus,

$$\mathbb{E}\left[\xi^2\mathbb{I}(\xi > a)\right] \leqslant 2\sqrt{2}\|\xi\|_{\psi_2}^2 \exp\left\{-\frac{a^2}{2\|\xi\|_{\psi_2}^2}\right\}.$$

Taking $a = \|\xi\|_{\psi_2}\sqrt{2\ln(4L\sqrt{2})}$, we finally obtain that

$$\mathbb{E}\left[\xi^2\mathbb{I}(\xi > a)\right] \leqslant \frac{\|\xi\|_{\psi_2}}{L} \leqslant \mathbb{E}\xi^2,$$

where the last inequality is due to the sub-Gaussianity of the class $\mathcal{F}$. Hence,

$$\mathbb{E}\left[\xi^2 e^{-\xi}\right] \geqslant \frac{e^{-a}}{2}\mathbb{E}\xi^2 = \frac{\mathbb{E}\xi^2}{2}\exp\left\{-\|\xi\|_{\psi_2}\sqrt{2\ln(4L\sqrt{2})}\right\} = \kappa_\xi\,\mathbb{E}\xi^2,$$

where we introduced

$$\kappa_\xi = \frac{1}{2}\exp\left\{-\|\xi\|_{\psi_2}\sqrt{2\ln(4L\sqrt{2})}\right\}.$$

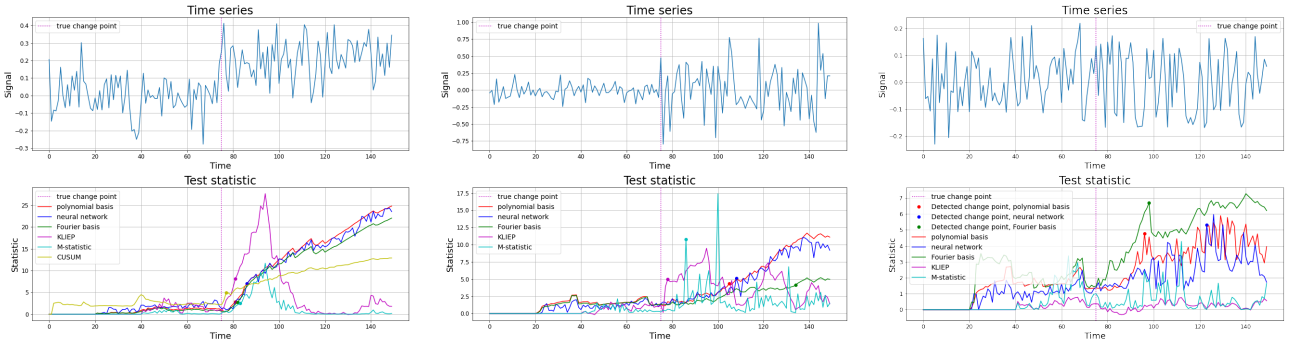
In other words, for any $f \in \mathcal{F}$, it holds that

$$\begin{aligned}
\kappa_{f(X_1)}\mathbb{E}f^2(X_1) &\leqslant \mathbb{E}f^2(X_1)e^{-|f(X_1)|} \\
&\leqslant \mathbb{E}\left[-f(X_1) + 2\ln\left(\frac{1+e^{f(X_1)}}{2}\right)\right] \\
&= -\frac{\tau(t-\tau)}{t}\mathbb{E}\mathcal{T}_{\tau,t}(f).
\end{aligned}$$

This yields the desired result.

E NUMERICAL EXPERIMENTS

This section contains additional information about numerical experiments, described in Section 4. We made all the calculations using a desktop computer with 16 GB RAM and CPU Intel Core i5-4690, 3.5 GHz and a laptop with 16 GB RAM and CPU Apple M1. Figure 3 shows an example of change point detection on three synthetic data sets, introduced in Section 4. The plots with observations are presented in the top line of Figure 3. The bottom line shows the corresponding values of the test statistic with different choices of the base class $\mathcal{F}$. The experiments show that Algorithm 1 with the class $\mathcal{F}$, corresponding to polynomials (solid red line) and neural networks (solid blue line) detects a structural change better than if one takes $\mathcal{F}$ equal to the linear span of several elements of the Fourier basis. Note that the classes of polynomials and neural networks ensure a better behaviour of the test statistic $\mathcal{S}_t$ in the stationary regime, yielding lower values of the threshold $\mathfrak{z}$. The same remarks are also true in the experiments on the CENSREC-1-C data set. One can find some examples of change point detection on this data set in Figure 4. Tables 4 and 5 contain the information about the thresholds and the parameters of algorithms in the experiments on the synthetic data and CENSREC-1-C. The thresholds and the values of hyperparameters in the experiments with the WISDM data set were already specified in Section 4. The plots of the test statistics in the experiments on this data set are displayed in Figure 5.

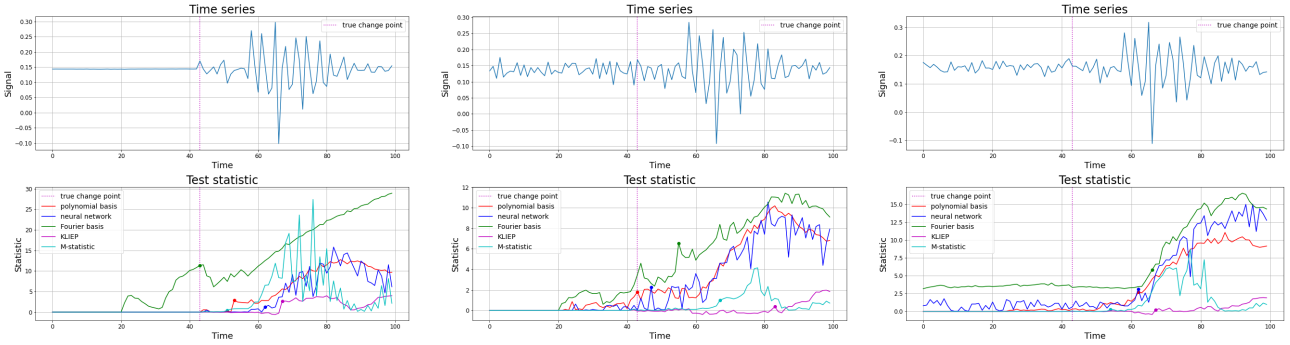


(a) (b) (c)

Figure 3: Examples of change point detection on synthetic data sets. Top line: the sequence of observations. Bottom line: corresponding values of the test statistic $\mathcal{S}_t$ with three variants of the base class $\mathcal{F}$: class of polynomials (red), linear span of several elements of Fourier basis (green) and class of neural networks (blue). Also, the bottom graph shows the statistics for the methods: CUSUM for mean shift detection (yellow), KLIEP (magenta) and M-statistics (cyan). The dashed vertical line corresponds to the true change point τ^*. The circle markers on solid lines correspond to the detection moments. Column (a): mean shift detection in a Gaussian sequence model (Example 1). Column (b): variance change detection in a Gaussian sequence model (Example 2). Column (c): distributional change from uniform distribution on $[-\sigma\sqrt{3}, \sigma\sqrt{3}]$, to the $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.1$.

Table 4: The thresholds $\mathfrak{z}$ and the values of hyperparameters of Algorithm 1 with different classes $\mathcal{F}$, the kernel change point detector, and KLIEP on synthetic data sets.

METHOD	EXAMPLE 1		EXAMPLE 2		EXAMPLE 3	
	$\mathfrak{z}$	PARAMETER	$\mathfrak{z}$	PARAMETER	$\mathfrak{z}$	PARAMETER
Algorithm 1 + polynomials	2.46	$p = 1$	3.98	$p = 2$	4.04	$p = 5$
Algorithm 1 + Fourier basis	2.49	$q = 2$	3.95	$q = 3$	6.84	$q = 6$
Algorithm 1 + neural networks	4.69	-	4.69	-	4.33	-
KLIEP	6.09	$b = 0.2$	4.19	$b = 0.33$	0.93	$b = 0.5$
M-statistic	9.59	$b = 0.5$	36.75	$b = 0.1$	17.65	$b = 0.25$
CUSUM	0.45	-	-	-	-	-



(a) (b) (c)

Figure 4: Examples of change point detection on the CENSREC-1-C data set. Top line: the sequence of observations. Bottom line: corresponding values of the test statistic $\mathcal{S}_t$ with three variants of the base class $\mathcal{F}$: class of polynomials (red), linear span of Fourier basis (green) and class of neural networks (blue). Also, the bottom graph shows the statistics for the methods: KLIEP (magenta) and M-statistics (cyan). The dashed vertical line corresponds to the true change point τ^*. The circle markers on solid lines correspond to the detection moments. Column (a): clean speech record. Column (b): speech record corrupted with noise, SNR = 20. Column (c): speech record corrupted with noise, SNR = 15.

Table 5: The thresholds $\mathfrak{z}$ and the values of hyperparameters of Algorithm 1 with different classes $\mathcal{F}$, the kernel change point detector, and KLIEP on the CENSREC-1-C data set.

	MAH_clean		MAH_N1_SNR20		MAH_N1_SNR15	
	$\mathfrak{z}$	PARAMETER	$\mathfrak{z}$	PARAMETER	$\mathfrak{z}$	PARAMETER
Algorithm 1 + polynomials	1.18	$p = 9$	1.75	$p = 9$	2.47	$p = 9$
Algorithm 1 + Fourier basis	10.83	$q = 10$	5.27	$q = 10$	5.26	$q = 10$
Algorithm 1 + neural networks	0.83	-	1.80	-	2.44	-
KLIEP	0.04	$b = 0.13$	0.32	$b = 0.19$	0.12	$b = 0.2$
M-statistic	1.20	$b = 0.04$	2.34	$b = 0.15$	0.86	$b = 0.1$

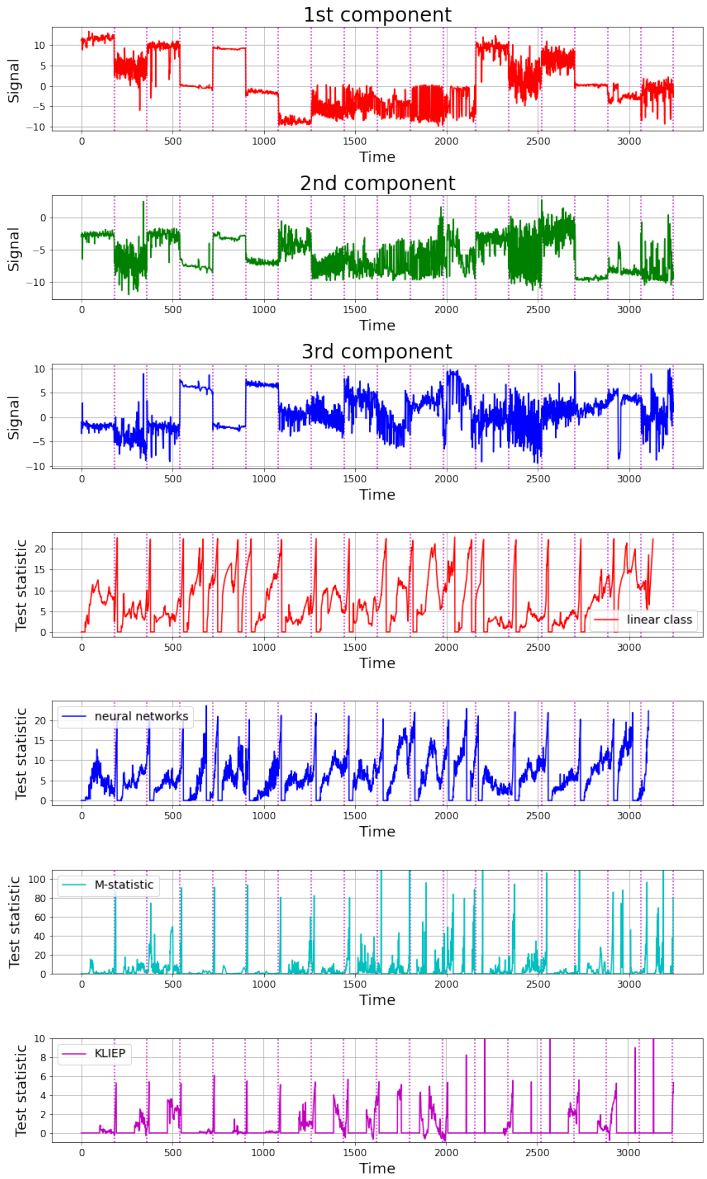


Figure 5: Three-dimensional time series from the WISDM data set and the corresponding values of the test statistics for Algorithm 1 (with two variants of the class $\mathcal{F}$, red and blue), the kernel change point detector with M-statistic (cyan) and KLIEP (magenta). The dotted vertical lines correspond to the moments of change points.