On Matched Filtering for Statistical Change Point Detection

Kevin C. Cheng, Shuchin Aeron, Michael C. Hughes, Eric L. Miller

Non-parametric and distribution-free two-sample tests have been the foundation of many change point detection algorithms. However, noise in the data make these tests susceptible to false positives and localization ambiguity. We address these issues by deriving asymptotically matched filters under standard IID assumptions on the data for various sliding window two-sample tests. In particular, in this paper we focus on the Wasserstein quantile test, the Wasserstein-1 distance test, maximum mean discrepancy (MMD) test, and the Kolmogorov-Smirnov (KS) test. To the best of our knowledge this is the first time an matched filtering has been proposed and evaluated for these tests or for change point detection. While in this paper we only consider a subset of tests, the proposed methodology and analysis can be extended to other tests. Quite remarkably, this simple post processing turns out to be quite robust in terms of mitigating false positives and improving change point localization, thereby making these distribution-free tests practically useful. We demonstrate this through experiments on synthetic data as well as activity recognition benchmarks. We further highlight and contrast several properties such as sensitivity of these tests and compare their relative performance.

Index Terms—Change point detection, matched filter, Wasserstein distance, Human activity data

I. INTRODUCTION

The foundational work by [1] and [2], later summarized by [3] formed the basis of many change point detection (CPD) methods today. Broadly classified, statistical CPD methods are either parametric, where changes are detected in the parameter space of some parametric model of the data which is either assumed apriori [4] or learned from data [5], or non-parametric, where the test statistic is derived directly from the samples.

Early work in parametric methods such as CUSUM detected change points by fitting some distributional model (e.g. Gaussian) to the data and developed tests to detect changes in the respective parameters. Another realm of parametric CPD methods consider ARMA-type dynamical models with state-space generalizations [6]. Generally, parametric methods are effective when the modelling assumptions hold and capture the data’s key characteristics.

More recently, there has been a growing interest in non-parametric CPD methods for applications where the characteristics marking the states or state transitions are unknown or cannot be modeled due to complexity or limited access to data. Applications to time series data are vast including human activity [7], ECG [8], EEG [9], speech signals [10], and climate data [11].

Classical non-parametric two-sample tests such as the Kolmogorov-Smirnov (KS), Cramer-von Mises, and Mann-Whitney statistics, have been applied to change point detection [12], [13]. Other methods use the family of f-divergences, such as the KL-divergence as a dissimilarity measure between empirical distributions. Recently, statistical two-sample tests belonging to the broader family of integral probability metrics [14] such as the maximum mean discrepancy (MMD) [15] and the Wasserstein-1 distance [16], have also been applied to change point detection [10].

These non-parametric change point methods often employ sliding windows of a fixed size to compute a test statistic that can be interpreted as a change point score as a function of time, where candidate change points are peaks of this function [17].

In tests that are distribution-free, the distribution of the test statistic under the null hypothesis that empirical windows are drawn from the same distribution is independent of data distribution. Therefore, threshold values that correspond to statistical confidence levels for change points can be applied universally.

In practice, basing change point decisions solely based on the local maxima of the two sample test can be insufficient to control false positives and difficulty with precise localization of a change point. In the current literature, this problem is avoided by considering only single change points [10] or using metrics such as AUC [18].

Our goal is to overcome these concerns while providing a unifying framework for understanding non-parametric distribution-free tests for change point detection. Our main contributions are:

- We propose the application of matched filtering to CPD to improve detection and localization of change points. In algorithms that employ sliding windows, the effects of a change point on the test statistic will be reflected in an interval around the change point. Therefore, we can derive a asymptotically matched filter from this expected response on this interval.
- We derive the asymptotically matched filters for the Kolmogorov-Smirnov (KS), maximum mean discrepancy (MMD), Wasserstein quantile (WQT), and Wasserstein-1 distance (WDT-M1) tests.
- We propose the sliced Wasserstein quantile test (SWQT), an extension of the WQT (which only applies to univariate observations) to multivariate observations.
- Even though the IID assumptions under which our asymptotically filter is derived are generally not satisfied, we show improvements in change point localization and false positive reduction when applied to real world data.
- We provide practical insight into fundamental differences between two-sample and quantile-quantile (Q-Q) tests in terms of sensitivity to data transformations, suggesting how application-specific criteria may drive which test is preferred.

This research was supported by funding from Army Research Center Natick via Tufts Center for Applied Brain and Cognitive Sciences (CABCS) under ARM994. Shuchin Aeron was supported in part by NSF CAREER award.
The remainder of this paper is organized as follows. In Sec. II-A, we introduce the CPD problem and required technical background. In Sec. III-A, we motivate and outline the simple approach of matched filtered two-sample tests applied to CPD. In Sec. III-A, we state the main theorems for the asymptotically matched filters for WQT, MMD, KS, and WDT-M1 tests. In Sec. IV, we demonstrate the performance of our method on simulated data as well as the commonly used human activity data set (HASC) and dancing honeybees (Beedance). Finally, in Sec. V, we illustrate and discuss of main differences between Q-Q and two-sample tests.

II. CHANGE POINT DETECTION

A. Problem Statement

We are given time series data, $X[t] \in \mathbb{R}^d, t = 1, 2, \ldots$, with the following model.

1) The data consists of distinct time segments $[0, \tau_1], [\tau_1 + 1, \tau_2], \ldots, [\tau_k + 1, \tau_k + 1], \ldots$, with $\tau_1 < \tau_2 < \ldots$, such that within each time segment, $X[t], t \in [\tau_k + 1, \tau_k + 1]$ are IID samples from a fixed but unknown distribution.

2) The distributions in successive time segments are different but in general two non-adjacent segments can have samples from the same distribution.

The time points $\tau_1, \tau_2, \ldots$ are referred to as the change points (CP). Given these conditions, the problem of Change Point Detection (CPD) is to estimate $\tau_1, \tau_2, \ldots$ without any information and assumptions (priors or parametric models) on the number and location of CPs.

Note that this approach to CPD is fundamentally an unsupervised machine learning problem. Related supervised CPD problems exist when some training data sequences are provided with labeled segment boundaries.

The IID assumption governing data within a segment greatly simplifies some of our later theoretical analysis, though it may inevitably be unrealistic for most real datasets. However, we demonstrate that our matched filter methods can be successfully deployed with their desirable properties intact (e.g. few false positives) even on real data where this assumption may be violated.

B. Notation

Given a collection of iid samples $X = \{x_1, x_2, \ldots, x_n\}, x_i \sim P$ we denote the estimated empirical distribution as $P_n \triangleq \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq t}$, the quantile function $P_n^{-1}(t) \triangleq \inf \{x : P_n(x) \geq t\}$ and the Q-Q function $P_n(Q_m)$. Also, $\rightarrow_w$ denotes weak convergence or convergence in distribution.

C. Statistical CPD with Matched Filter

Our general framework to addressing this problem is to use statistical two-sample tests comparing adjacent sliding windows of a constant size as our test statistic for CPD.

Given time-series $X[t]$ on a compact domain $C \subset \mathbb{R}^d$, we define two empirical CDFs at each time $t \geq n$; one generated from the sum of dirac-delta functions supported on a window of size $n$ to the left of $t$, $F_n(t) \triangleq \{X[t-n], X[t-n+1], \ldots, X[t-1]\}$ and the other from the $n$ samples to the right $G_n(t) \triangleq \{X[t+n], X[t+n-1], \ldots, X[t+1]\}$. A statistical test $D_n(F_n, G_n)$ is then applied to these windows. With a slight abuse of notation, we define the test statistic over the pair of sliding windows $D_n(t) = D_n(F_n(t), G_n(t))$.

The nominal approach for identifying change points given a test statistic is to label local maxima of a computed statistic above some threshold parameter $\theta$. However, as evident in Fig. 1 noise in the signal can push the statistic above the threshold causing false detections. Furthermore, in the presence of change points, multiple peaks ambiguities the exact localization of change points.

In signal processing detection and estimation, these issues are address by deriving a matched filter, commonly known as the time flipped conjugate version of the signal to be detected, which is statistically optimal given additive white Gaussian noise (AWGN). While the AWGN does not apply in our problem, the motivation for signal detection still applies.

Since for two-sample tests, the test statistics can be interpreted as confidence levels for rejection the null hypothesis, the matched filtering process should be peak preserving. We call such a filter peak preserving if the expected value of the unfiltered is equal to filtered signal at an expected change point.

For a given matched filter $h[t]$, the peak preserving constant $\alpha = (\sum_t h[t]^2)^{-1}$, the resulting filtered statistic is then:

$$F* = D_n[t] \oplus \alpha h[t],$$

where $\oplus$ denotes the convolution operation.

III. MATCHED FILTERS FOR STATISTICAL TESTS

In this section we outline the general framework used to derive the asymptotically matched filters and state the main results for the various statistical tests. Interested readers can find the complete proofs in the supplemental material.

In the following we will assume that $X = \{X_1, X_2, \ldots, X_n\}$ are IID $\sim P$, and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ IID $\sim Q$. When they are supported on $\mathbb{R}$ their respective empirical measures are denoted by $P_n$ and $Q_m$ respectively.

Generally, two-sample tests are applied by directly computing the distance between measures. Q-Q tests are a subset of two-sample tests that compute the distance between the Q-Q

Footnote

Here we use $D_n$ as a general statistical test and it can be substituted for the various statistical tests defined in Sec. III-A.
Wasserstein-p distance, or earth mover’s distance, is defined as applied to any statistical test under this framework. For a test statistic converges to a deterministic function of \( \pi \), whereas the distribution of \( \tau \) remains constant, the resulting matched filter for when \( n \) is sufficiently large (but finite) is linear respect to \( \pi_1 \), thus \( h(t) = (1 - \frac{|t|}{\pi}) \).

**C. Wasserstein Quantile Two-Sample Test**

The Wasserstein Quantile Two-Sample Test (WQT) is a distribution-free variant of the WDT that measures the Wasserstein distance of the Q-Q function of \( P_n, Q_m \) to the uniform measure \( [22] \).

For finite \( n \), when using the WQT, the peak preserving matched filter (to be applied as stated in (1)) can be approximated by \( h(t) = (1 - \frac{|t|}{\pi})^2 \).

For the WQT, the theorem is stated with \( d \) being 1, which essentially removes a distribution dependent \( O(1) \) term in the WQT. In the operational case, we approximate this term to be \( \int_0^1 E^B(x) dx \), where \( B(x) \) is the standard Brownian bridge on \( [0, 1] \). This acts as a constant bias term that is removed from the signal prior to peak matched filter convolution when using the WQT. The details of this can be found in the supplementary material.

**D. Sliced Wasserstein Quantile Test**

Since the WQT in only defined in one dimension, the naive approach for extension to multiple dimensions to average the WQT across each dimension independently. Alternatively, we can use the approach taken for the sliced Wasserstein distance \([24]\). Formally this is

\[
S_{\text{swqt}}(F_n, G_m) = \int_{S^{d-1}} D_{\text{WQT}}(F_n^\theta, G_m^\theta) d\theta
\]

Where \( d\theta \) is a uniform measure on \( S^{d-1} \), the unit sphere in \( \mathbb{R}^d \), and \( F_n^\theta \), \( G_n^\theta \) are the respective CDFs given the \( \mathbb{R}^d \to \mathbb{R}^d \) projection of the samples according the unit vector \( \theta \).

With this definition we state the following.

**Theorem 3:** Given the setup and assumptions in Sec. III-A and \( d_{\text{swqt}}(P, Q) = \int_{S^{d-1}} d_{\text{swqt}}(P^\theta, Q^\theta) d\theta \),

\[
\frac{1}{n} D_{\text{swqt}}(P_n, Q_m) \rightarrow \pi_1^2 d_{\text{swqt}}(P, Q)
\]
We note the following theorem for the KS test. When this kernel is bounded, the MMD is known to be

\[ MMD = \sup x \in \mathbb{R}^d |P_n(x) - Q_n(x)| \]

With a simple change of variables \( y = G_n(x) \) we can also show that the KS test is a Q-Q test [26].

\[ D_{KS}(F_n, G_n) = \sup_y |F_n(G_n^{-1}(y)) - y| \]

We note the following theorem for the KS test.

**Theorem. 4:** Given the setup and assumptions in Sec. III-A and \( d_{KS}(P, Q) = \sup_y |P - Q| \)

\[ D_{KS} \rightarrow_{w} \pi_1 d_{KS}(P, Q) \]

Under the null hypothesis \( D_{KS} \rightarrow 0 \), the peak preserving matched filter for the the KS test can be approximated (to be applied as stated in [1]) by \( h[t] = (1 - |t|^2) \).

**E. Kolmogorov-Smirnov**

The two-sample KS [25] test computes the maximum deviation between respective empirical distribution functions

\[ D_{KS}(P_n, Q_n) = \sup x |P_n(x) - Q_n(x)| \]

(9)

With a simple change of variables \( y = G_n(x) \) we can also show that the KS test is a Q-Q test [26].

\[ D_{KS}(F_n, G_n) = \sup_y |F_n(G_n^{-1}(y)) - y| \]

(10)

We note the following theorem for the KS test.

**Theorem. 4:** Given the setup and assumptions in Sec. III-A and \( d_{KS}(P, Q) = \sup_y |P - Q| \)

\[ D_{KS} \rightarrow_{w} \pi_1 d_{KS}(P, Q) \]

(11)

Under the null hypothesis \( D_{KS} \rightarrow 0 \), the peak preserving matched filter for the the KS test can be approximated (to be applied as stated in [1]) by \( h[t] = (1 - |t|^2) \).

**F. Maximum Mean Discrepancy**

The MMD represents the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS) with kernel \( k(x, y) \).

\[ \text{MMD}^2(k, P, Q) = \sup_{f \in \text{RKHS}(k)} \|f\|_k \leq 1 \left( E_P[f] - E_Q[f] \right)^2 \]

When this kernel is bounded, the MMD is known to be distribution-free. [13].

For this case we note the following theorem.

**Theorem. 5:** Given the setup assumptions in Sec. III-A and \( d_{mmd}(P, Q) = E_P \times P \times [k(x_1, x_2)] + E_Q \times Q \times [k(y_1, y_2)] - 2E_P \times Q [k(x_1, y_2)] \)

\[ E[\text{MMD}^2(P_n, Q_m, k)] = \pi_1^2 d_{mmd}(P, Q) \]

(12)

Given that under the null hypothesis, \( E[D_{mmd}] = 0 \), the peak preserving matched filter (to be applied as stated in [1]) has \( h[t] = (1 - |t|^2), \mu = 0 \)

IV. Evaluation

A. Simulation Data

We first verify our proposed matched filters on simulated data. Given two known distributions \( P \) and \( Q \), we simulate exactly the mixture scenario in Fig. 2 where \( F_n \sim \pi_1 P + (1 - \pi_1)Q \) and \( G_n \sim Q \). We vary the mixture proportion \( \pi_1 \) and compute the test statistic \( D_n(F_n, G_n) \) for various window sizes \( n \). Each test was averaged over 100 repetitions using different random seeds.

The performance of the matched filters are evaluated in for the univariate (for the tests that are only defined on univariate data), and multivariate case.

For the univariate case, we generate 40 IID data sequences of length 800 with a single change point randomly distributed between [300,500]. Samples prior to the change point are drawn from distribution \( P \sim \mathcal{N}(0, 1) \), whereas samples after the change point are drawn from \( Q \sim \mathcal{N}(0.25, 1) \). Change points are detected as peaks that exceed a threshold with significance level of 5%.

For the multivariate case, when we observe 2-dimensional data vectors again simulating 40 sequences of length 800 with a single change point between [300, 500]. We define a common covariance matrix \( \Sigma \) with unit diagonal and high correlation (0.9), and then define \( P \) and \( Q \) so they differ in mean:

\[ P \sim \mathcal{N}\left( \begin{bmatrix} -0.12 \\ 0.12 \end{bmatrix}, \Sigma \right) \]

\[ Q \sim \mathcal{N}\left( \begin{bmatrix} +0.12 \\ -0.12 \end{bmatrix}, \Sigma \right) \]

\[ \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \]

With these datasets, we compare the performance between the filtered and unfiltered test statistics for various window lengths. Change point localization is measured by the area under the precision recall curve (AU-PRC), and false positives are measured by the average number peaks above the detection threshold. For a given window size \( n \) peaks are labeled as true positives if they fall within \( n \) samples of a true change point. To be fair to the unfiltered tests, duplicate peaks within \( n \) samples are ignored.

To validate the performance of the sliced Wasserstein quantile test (SWQT) as an appropriate extension of the WQT to higher dimensions, we compare it to the naive approach of simply averaging WQT over each dimension independently. The SWQT is computed via Monte Carlo simulations by randomly sampling vectors \( \theta \sim S^{d-1} \), and averaging the results over each linear project. For the MMD we define the kernel \( k(x, y) \) as the Gaussian kernel with unit variance.
**B. Simulated Results**

Fig. 3 plots the value of the average test statistic as a function of \( \pi_1 \) for \( P \) and \( Q \). These plots confirm the results from our theorems and show the convergence of each of the statistical tests to the expected distribution fit.

In the simulated change point tests on \( \mathbb{R}^1 \) (Tab. II), and \( \mathbb{R}^2 \) (Tab. III) we see that applying our proposed matched filters to the corresponding test statistic yields consistent improvement in both change point localization (as shown by higher AU-PRC value), as well as a decrease in the false detection rate. As expected, when window size increases overall detection performance also increases.

The results also show that when extending to multiple dimensions, the SQT greatly outperforms the results from averaging the WQT over independent dimensions, demonstrating the SWQT is a suitable extension of the WQT to higher dimensions.

In this controlled setting, the performance across all four possible statistical tests is relatively consistent in the univariate case. In contrast, in the two-dimensional case the SWQT has a better overall performance compared to the MMD across all window sizes.

We also highlight that since the data matches all of our assumptions, we are also able to verify the peak preserving property of the matched filter (Fig. 1).

**C. Real World Data**

The only hyperparameters required for the non-parametric statistical change point methods described in this paper are the window size \( n \) and detection threshold parameter. Since the selection of threshold value varies between tests, we choose to show the precision-recall plots on the real world datasets to allow assessment of performance across possible thresholds. Given domain knowledge of the dataset, we choose the window size \( n \) based on the scale at which change points should be detected, and a detection window \( \delta \) which defines the maximum
distance to ground truth for change points to be labeled true positives.

To compare effectively to prior work, we also report area under the receiver operator curve (AU-ROC). The AU-ROC is computed purely based on the test statistic scores and ground truth change point labels. Notably, here we do not apply the AU-ROC as defined in [17] as the monotonicity guaranteed. Rather we follow the AU-ROC computation as described in [18]. Ultimately, if we consider the problem of multiple CPD as a binary classification problem (CP vs no CP), the AU-ROC is a poor metric for assessing classification performance under severe class imbalance, as we have in CPD where only a small fraction of timesteps will generally be true change points. We recommend using the precision-recall curve (and the AU-PRC) instead, but report AU-ROC to engage with prior work.

We compare the filtered and unfiltered versions of the statistical tests described in this paper with the M-statistic [7], an unsupervised MMD based sliding window CPD algorithm to which we apply identical windowing parameters. Under the AU-ROC metric we also compare three supervised CPD methods. First, KL-CPD [18] applies the MMD with a kernel trained as a neural network. Next, KLIEP [27] and RuLSIF [7] use density ratio tests as the basis for CPD.

We evaluate on the following datasets:

HASC-PAC2016: [19] consists of over 700 three-axis accelerometer sequences of subjects performing six actions: ‘stay’, ‘walk’, ‘jog’, ‘skip’, ‘stairs up’, and ‘stairs down’. We evaluate on the 100 longest sequences where each of the actions are represented. A window size of 200 samples (2 seconds) and a detection window of \( \delta = \pm 150 \) samples were chosen.

Beedance: [20] movements of dancing honeybees who communicate through three actions: “turn left”, “turn right” and “waggle”. The data, taken from images uses X, Y pixel location as well as angle \( \theta \) of the bee. We use the gradient of the original data, with a window size of 12 samples, and detection window of \( \delta = \pm 10 \) samples.

Given that both datasets belong in \( \mathbb{R}^3 \), methods inherently defined in \( \mathbb{R}^3 \) are extended to higher dimensions by averaging their respective test statistic over each dimension.

D. Real World Results

In Tab. IV our reported AU-ROC on Beedance for the WDT-M1 tests exceeds that the other proposed methods. On HASC, our unsupervised methods are well within the range of reported AU-ROC metrics even for supervised methods.

Fig. 4 shows the precision-recall precision curves for both the unfiltered and filtered change point statistics across both datasets. Since only peaks of the statistic are considered, this is not a standard binary hypothesis test and therefore the precision-recall may not necessarily always reach 100% recall. For fairer comparisons we did not count duplicate peaks within the detection window in the unfiltered statistic. Consistent with the simulated results we see that with the matched filter applied, all tests show an improvement in precision which suggests better change point localization.

Furthermore, since there is only a slight improvement in the SWQT versus the WQT in both HASC and Beedance, this is indicative that change points observed by the WQT can be observed in the independent dimensions.

The Beedance dataset is a test of the asymptotic properties of our statistical tests. Due to the low sampling rate in proportion to the frequency of change points, the selected window size of 12 samples negatively impacts the KS and WQT tests the most.

Especially in the HASC precision-recall curve, we see that the performance of the statistical tests reported vary widely. WDT-M1 and MMD have the highest precision, KS is a middle ground, and the WQT and SWQT generally have the lowest precision for a given recall. This is indicative of that the Q-Q tests on the HASC data false alarm more often compare to the two-sample tests (Fig. 5). Following this we can learn how Q-Q tests behave differently from two-sample test and when we can leverage those differences.

V. QUANTILE-QUANTILE VS TWO-SAMPLE TESTS

These differences can be traced to the structure of Q-Q tests versus two-sample tests. Fundamentally, Q-Q tests will be invariant to transformations where the Q-Q map stays constant. For example if we scale two distributions with the same factor, the resulting Q-Q map will not change and thus the statistic under a Q-Q test will also stay constant. Furthermore, since Q-Q maps are monotone and bounded from \([0, 1]\) the resulting testing statistic will also be bounded. For example for a finite window size, WQT has a maximum value, which is achieved when the 0-th quantile of \( P \) maps to the 100-th quantile of \( Q \) (see supplement). Conversely, two-sample tests like MMD and WDT-M1 are not bounded and will scale as a function of distance between the distributions.

This highlights the different regions of sensitivity of Q-Q and two-sample tests. We see this dichotomy clearly in the WQT and WDT-M1 (Fig. 6) when comparing uniform distributions.
Fig. 5. Sample output for HASC-PAC2016 human activity accelerometer data sequence (grey) with the filtered (solid) and unfiltered (dashed) SWQT (blue), and MW1 (purple). For comparison, the SWQT and MW1 are normalized based on their peak value over the sequence. While it appears that the WQT false alarms frequently, the right insert shows a small shift in mean that is detected by WQT but not WDT-M1.

Fig. 6. Comparison of sensitivity for Q-Q (red) and 2-sample (purple) tests. We compare two uniform distributions of shifting supports (left) as well as two normal distributions with varying mean and variance parameters (right). In both cases the WQT has higher sensitivity to small changes in support but quickly saturates whereas the WDT scales indefinitely as the distributions diverge.

with shifting supports and normal distributions with shifting means and variances. The WQT is highly sensitive to small changes, especially changes in the mean parameter, but also quickly saturates whereas the sensitivity of the WDT remains constant in all regions.

Fig. 5 shows an example of these differences in human activity data. The different regions of sensitivity are highlighted in the fact that the WQT generally has peaks of equal height whereas the peaks of the WDT-M1 scale with the observed magnitude of change. Furthermore, while it appears that the WQT false alarms in the stationary regions, closer inspection into one of these regions shows that the signal has a slight shift in mean. Since the signal is relatively stable in those regions, the shift causes a change in support which triggers the WQT to alarm. This change is not reflected in the WDT-M1 because the scale of the change is dwarfed in scale by the other changes in the signal. In the HASC data, these small changes are not labeled as change points, which contributes to the poor precision of the Q-Q tests. However, whether or not such subtle shifts in mean should be flagged as true change points is application specific. Depending on the desired behavior, a practitioner may prefer either Q-Q tests (more sensitive to subtle shifts) or 2-sample tests (less sensitive).

VI. CONCLUSION AND FUTURE WORK

While many methods of change point detection have been proposed over the years, the issue of change point localization for a noisy distribution-free statistic not been properly addressed. To overcome this issue we introduce asymptotically matched filters. We derive these filters for various non-parameteric two-sample and Q-Q tests, that have been used as the foundation of many CPD algorithms, under the simple setup that sliding windows over a change point will cause samples from one window to be drawn from a mixture distribution. Clearly this analysis can be applied beyond the statistical tests discussed in this work.

While simple in concept, the advantages of change point detection under this framework is that there are the very limited amount of hyperparameters required for model selection (window size and detection threshold), and the statistical guarantees that come with the analysis.

Despite the fact that the derivation for the filters in this paper assume that the data is IID, based on real world results, we see that the benefits still hold on non-IID data. Nonetheless, in future work, we hope to consider analysis under non-IID conditions, and the derivation of matched filters for other change point tests.

Lastly, we illustrate some fundamental differences in sensitivity between two-sample tests versus Q-Q tests. These insights could be leveraged to properly select the appropriate test for an application, and certainly motivate further rigorous investigation.

APPENDIX

In this supplement we prove the individual theorems that are used to derive the asymptotically matched filters for two-sample change point detection by computing the expected value of the two-sample test in the region around a change point in the limit as the window length $n$ goes to infinity. We use these asymptotic results to design a peak-preserving filter that can be applied in the case where $n$ is finite.

In Sec. A we present the problem setup and assumptions, and in Sec. B we provide some mathematical background. We then prove Thm. 1 from the main paper for the Wasserstein-1 distance test with Minkowski metric (WDT-M1) in Sec. C.
Thm. 2 for the Wasserstein quantile test (WQT) in Sec. D, Thm. 3 for the sliced Wasserstein quantile test (SWQT) in Sec. A, and Thm. 4 for Kolmogorov-Smirnov (KS) in Sec. F, and Thm. 5 for the maximum mean discrepancy (MMD) in Sec. G.

A. Derivation Assumptions and Setup

We consider a discrete-time, stochastic sequence \( X[t] \) with a change point at \( t = \tau \), such that samples are drawn IID from a distribution \( P \) for \( (\tau - 2n) \leq t < \tau \), and from another distribution \( Q \) for \( \tau \leq t < (\tau + 2n) \) (Fig. 7). This requirement for the \( 4n \) samples around \( \tau \) stems from the sliding window framework, where the change point will be included in either \( F_n[t] \) or \( G_n[t] \) for \( (\tau - n) \leq t < (\tau + n) \), thus drawing from samples for \( (\tau - 2n) \leq t < (\tau + 2n) \). We also assume distributions \( P \) and \( Q \) are continuous and are supported on a compact domain \( C \).

At each point, we define two empirical measures derived from the \( n \) points prior to and after \( t \). That is, the \( n \) samples to the left of \( t \) are used for \( F_n[t] = \{ X[t-n], X[t-n+1],...,X[t-1] \} \), and the \( n \) samples to the right of \( t \) generate \( G_n[t] = \{ X[t], X[t+1],...,X[t+n-1] \} \). A statistical two-sample test \( D_\pi[t] = D_\pi(F_n[t],G_n[t]) \) is then applied to this pair of sliding windows.

Without loss of generality (as long as we show \( D_\pi \) to be symmetric), we consider the case where we start at \( t = \tau \) and slide the window to the right such that the change point is located in the \( left \) set of samples. In this setup, the left window will be drawn from a mixture distribution \( F_n[t] \sim (\pi_1 P + (1-\pi_1) Q) \) with mixture proportion \( \pi_1 = \left(1 - \frac{t+\tau}{n}\right) \) as we slide across the interval from \( t \in [\tau, \tau + n] \), which corresponds to \( \pi_1 \in [0, 1] \). The distribution of the right window remains constant, \( G_n[t] \sim Q \). With a slight abuse of notation, we reference these windows as \( F_n, G_n \), where \( t \) is implied from the mixture proportion \( \pi_1 \).

In signal detection theory, a matched filter is statistically optimal given white Gaussian noise, and is commonly known to be the time reversed copy of the signal to be recognized [21]. While AWGN assumption does not apply in this case, the motivation for matched filtering to improve signal detection still applies.

Therefore, after showing that each two-sample test \( D_\pi \) is symmetric, the asymptotic matched filter is equal to \( \mathbb{E}[D_\pi(F_n,G_n)] \) as \( n \to \infty \) for \( \pi_1 \in [0, 1] \). This is the main result for each of our theorems that form the basis for the matched filter.

Furthermore, we show that these filters have the form,

\[
\tilde{h}(\pi_1, P, Q) = d_\pi(P, Q)h(\pi_1)
\]

where we normalize \( h(\pi_1) \) such that \( h(\pi_1 = 1) = 1 \) (corresponding to \( t = \tau \)). We observe that only the scale of the response \( d_\pi(P, Q) \) depends on the distributions \( P, Q \), and the time dependent component \( h(\pi_1) \), is distribution independent. These properties are necessary for the resulting operational matched filter to be distribution-free and peak-preserving. We define a filter to be peak-preserving if for \( g(t) = \alpha h(t), g(0) = (g(t) \circ h(t))(0) \). We apply these results in the asymptotic case for the continuous function \( h(\pi_1) \) for the operational case when \( n \) is finite \( h[t] \). With a slight abuse of notation,

Asymptotic: \( h(\pi_1) = h\left(1 - \frac{t-\tau}{n}\right), \pi_1 \in [0, 1] \)

Operational: \( h[t - \tau] = h\left(1 - \frac{t-\tau}{n}\right), t = (\tau - n), \ldots, (\tau + n) \).

Therefore, we can define the discrete-time matched filter \( h[t] \) as,

\[
h[t] = \begin{cases} 
  h \left(1 - \frac{|t|}{n}\right) & t = -n, (-n + 1), \ldots, n \\
  0 & else 
\end{cases}
\]

As described in the main paper (Eq. 1), with a peak-preserving scale factor \( \alpha \left(\sum_{t \in [-n,n]} h[t]^2\right)^{-1} \), the matched filtered test statistic is computed by,

\[
F_\pi[t] = D_\pi[t] \circ \alpha h[t]
\]

To prove the peak-preserving property, we know that from (13) and (14),

\[
\mathbb{E}[D_\pi[t = \tau]] = d_\pi(P, Q).
\]

Therefore, the filtered test statistic at \( t = \tau \) as defined in (16) is,

\[
\mathbb{E}[F_\pi[t = \tau]] = d_\pi(P, Q) \alpha \sum_{t \in [-n,n]} h[t]h[-t] = d_\pi(P, Q)
\]

since \( h[t] \) is symmetric.

B. Notation and Mathematical Preliminaries

Given probability distribution functions \( P \) and \( Q \), we use the following definitions for the empirical distribution function [19], quantile function [20], and quantile-quantile function [21].
\[ P_n(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} \quad C \to [0, 1] \]  
\[ P_n^{-1}(t) \triangleq \inf \{ x : P_n(x) \geq t \} \quad [0, 1] \to C \]  
\[ P_n(Q_m^{-1})(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq \inf \{ x : Q_m(x) \geq t \}} \quad [0, 1] \to [0, 1]. \]

Theorem A.1: [2.7 from [30]]

Theorem: A.1: \( \mathbb{1}_X \) vector space under the Skohorod metric (**Verify

\[ \text{for discontinuities in } X \text{ belong to a measurable metric space } S \text{ theorem.} \]

Given a mapping \( P \) to other domains, we utilize the continuous mapping of stochastic processes on their respective domains.

Consider a sequence of probability distributions over a set of continuous real functions \( P \) where \( \text{a Borel } \sigma \text{-field} \text{ (of all c `adl`ag functions over a compact domain, (} \) \text{ we can assign a measure } S \text{ convergence on metric spaces which we can apply here. The process over } X \text{ signifies weak convergence or convergence in distribution of a } \text{continuous real functions } \text{ where } \] \( \text{is a sequence of probability measures on } \text{is the standard Brownian bridge on } [0, 1], \text{ and } \text{for empirical distributions } P_n \text{ and } Q_m \text{ of } n, m \text{ samples respectively, where } \frac{n}{m} = \lambda \in [0, \infty) \text{ as } n, m \to \infty, \]

\[ \sqrt{\frac{nm}{n+m}} (G_m(F_n^{-1}(\cdot)) - G(F^{-1}(\cdot))) \to \]  
\[ \sqrt{\frac{\lambda}{\lambda+1}} B_1(G \cdot F^{-1}(\cdot)) + \sqrt{\frac{1}{1+\lambda f(F^{-1}(\cdot))}} B_2(\cdot) \]  

where \( B_1(x), B_2(x) \) are independent Brownian bridges .

To show the symmetric property of the Wasserstein quantile test, we also use the inverse function theorem.

Theorem A.4: [From [22]], Inverse Function Theorem: Suppose that \( r : [a, b] \to [r(a), r(b)] \) is monotone and continuous, then

\[ \int_{r(a)}^{r(b)} r^{-1}(y) dy = br(b) - ar(a) - \int_{a}^{b} r(x) dx. \]

Lastly, we state the Glivenko-Cantelli theorem often used in conjunction with the Kolmogorov-Smirnov test. Here, \( \to \) denotes almost-sure convergence which is a stronger condition than weak convergence.

Theorem A.5: [19.1 from [28]], Glivenko-Cantelli Theorem: If \( X_1, X_2, \ldots \) are IID random variables with distribution function \( F \), then \( \| F_n - F \|_{\infty} \to \) 0.

With this background we prove the asymptotic results for our various statistical two-sample tests.

C. Proof for Wasserstein-1 Distance in \( \mathbb{R} \) with Minkowski metric

Here we prove Thm. 1 for the asymptotically matched filter for the Wasserstein-1 distance test with Minkowski metric.

Theorem 1: Given the setup and assumptions in Sec. A and constant \( d_{\text{wtdt-1}}(P, Q) = \int \left| \left( P(x) - Q(x) \right) \right| dx \)

\[ D_{\text{wtdt-1}}(F_n, G_n) \to \pi d_{\text{wtdt-1}}(P, Q). \]

Proof. The WDT-M1 can be expressed as,

\[ D_{\text{d}}(F_n, G_m) = \int_c \left| F_n(x) - G_n(x) \right| dx. \]

We expand \( A(x) \) as,

\[ A_n = (F_n - F) - (G_n - G) + (F - G). \]

\[ \text{the full statement of the other conditions can be found in [30]}. \]
By the Glivenko-Cantelli theorem, (Thm. A.5)
\[
\sup_x |A_n(x) - (F(x) - G(x))| \to_{as} 0 \quad (28)
\]
\[
|A_n| \to_{as} |(F - G)| = |\pi_1(P - Q)| \quad (29)
\]
Since almost sure convergence implies weak convergence, we can apply the continuous mapping theorem for the function \(f(x)dx\) (Thm. A.1)
\[
D_{MW1}(F_n, G_m) = \int_C |A_n(x)|dx \to_w \pi_1 \int_C |(P(x) - Q(x))|dx = \pi_1 d_{wdt-m1}(P, Q). \quad (30)
\]

Given that the sequence \(\sqrt{\frac{2}{n}} \to 0\), we apply Slutsky’s theorem to \(35\),
\[
A_n \to_w 0_n \quad (36)
\]
\[
A_n + B_n \to_w F(G^{-1}(x)) - x. \quad (37)
\]

Therefore, by the continuous mapping theorem for the continuous function \(\int_0^1 (f(x))^2dx,\)
\[
\frac{1}{n} D_{wqt}(F_n, G_n) = \frac{1}{2} \int_0^1 (A(x) + B(x))^2 dx \to_w \frac{1}{2} \int_0^1 (F(G^{-1}(x) - x))^2 dx \quad (38)
\]
Lastly, since \(F = \pi_1P + (1 - \pi_1)Q,\) and \(G = Q,\)
\[
\frac{1}{n} D_{wqt}(F_n, G_n) \to_w \frac{1}{2} \int_0^1 (((\pi_1 P + (1 - \pi_1) Q) \circ Q^{-1})(x) - x)^2 dx = \frac{1}{2} \int_0^1 (\pi_1{(P}{Q^{-1}(x))} + {Q}{Q^{-1}(x)} - \pi_1{Q}{Q^{-1}(x)} - x)^2 dx = \pi_1^2 \int_0^1 (P{Q^{-1}(x)} - x)^2 dx = \pi_1^2 d_{wqt}(P, Q). \quad (39)
\]

Once again, since the WQT is bounded, assuming that \(F_n, G_n : C \to [0, 1],\) by Thm. 2 and the Portmaneau theorem, it follows that,
\[
E\left[\frac{1}{n} D_{wqt}(F_n, G_n)\right] = \pi_1^2 d_{wqt}(P, Q). \quad (40)
\]

This suggests that for a finite window size \(n,\)
\[
E[D_{wqt}(F_n, G_n)] = n \pi_1^2 d_{wqt} + O(1). \quad (41)
\]

For a given \(P\) and \(Q,\) \(B(x)\) is deterministic. To \(A_n\) we apply Thm. A.3
\[
\sqrt{n} A_n(\cdot) \to_w \sqrt{\frac{1}{2}} B_1(G \circ F^{-1}(-)) + \sqrt{\frac{1}{2}} g(F^{-1}(\cdot)) B_2(\cdot). \quad (35)
\]

To prove the symmetric property of the Wasserstein two-sample test we follow the same steps as \(34\) to \(38\) when the arguments are swapped to show that \(\frac{1}{n} D_{wqt}(G_n, F_n) \to_w\)
The symmetric property of the KS tests is easily seen since \( \frac{1}{n} D_{swqt}(F_n, G_n) \rightarrow_w \pi_1 d_{swqt}(P, Q) \). Here we use the same reasoning as in Sec. [4] for the operational case where the window size \( n \) is finite, 
\[
\mathbb{E}[D_{swqt}(F_n, G_n)] = n \pi_1^2 d_{swqt}(P, Q) + O(1).
\]
We estimate the \( O(1) \) term by considering the expected value of the WQT under the null hypothesis that \( P = Q \). In this case, 
\[
\mathbb{E}[d_{swqt}(P, Q)] = \mu_{\mathbb{B}_2} \text{ for all } \theta.
\]
Therefore, the expected asymptotic behavior of the SWQT can be approximated as follows,
\[
\mathbb{E}[D_{swqt}(F_n, G_n)] \approx n \pi_1^2 d_{swqt} + \mu_{\mathbb{B}_2}.
\]

### F. Proof for Kolmogorov-Smirnov Test

In this section, we compute the expected response of the Kolmogorov-Smirnov test (KS) around the change point by analysis of its asymptotic behavior as a function of mixture proportion.

**Theorem 4:** Given the setup and assumptions in Sec. [4] and constant 
\( d_{swqt}(P, Q) = \int_{S^{d-1}} d_{swqt}(P^\theta, Q^\theta) d\theta \),
\[
D_{KS}(F_n, G_n) \rightarrow_w \pi_1 d_{KS}(P, Q).
\]

**Proof.** The two-sample KS test is defined by:
\[
\pi_1 d_{KS}(P, Q).
\]

By the Glivenko-Cantelli theorem (Thm. [4.3]),
\[
D_{KS}(F_n, G_n) \rightarrow_{as} \sup \{ P(\theta) - Q(\theta) \}.
\]

Then, since \( F = \pi_1 P + (1 - \pi_1)Q \) and \( G = Q \),
\[
D_{KS}(F_n, G_n) \rightarrow_{as} \pi_1 \sup \{ P(\theta) - Q(\theta) \} = \pi_1 d_{KS}(P, Q).
\]

Therefore, it follows that the expected value of the KS tests converges to,
\[
\mathbb{E}[D_{KS}(F_n, G_n)] \rightarrow \pi_1 d_{KS}(P, Q)
\]

The symmetric property of the KS tests is easily seen since 
\( |P_n(x) - Q_n(x)| = |Q_n(x) - P_n(x)| \). We note that the matched filter for the KS test matches that of the WDT-M1.
we decompose the empirical MMD distance of the mixture $D_{\{Q\}}$.

Since it is defined in $\mathbb{R}^d$, the MMD is computed between the two sets of samples rather than their respective empirical measures.

**Theorem 5:** Given the setup and assumptions in Sec. A where $\{F_n\} = \{X[t - n], \ldots, X[t - 1]\}$, $\{G_n\} = \{X[t], \ldots, X[t + n - 1]\}$, and $d_{mmd}(P, Q) = \mathbb{E}_{P \times P}[k(p_i, p_j)] + \mathbb{E}_{Q \times Q}[k(q_i, q_j)] - 2\mathbb{E}_{P \times Q}[k(p_i, q_j)]$, where $p_i \sim P$ and $q_i \sim Q$ then,

$$E[D_{mmd}(\{F_n\}, \{G_n\}, k)] = \pi_1^2 d_{mmd}(P, Q).$$  \hspace{1cm} (55)

**Proof.** Given a symmetric kernel, $k(x, y)$, the MMD distance between two samples $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ is

$$MMD^2(X, Y) = \frac{1}{n^2 - n}\left( \sum_{i=1}^{n} k(x_i, x_j) - 2 \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} k(x_i, y_j) + \sum_{i=1}^{n} k(y_i, y_j) \right).$$

\hspace{1cm} (56)

Once again, for two windows of $n$ samples $F_n$ will have $\pi_1 n$ samples drawn from $P$ and $(n - \pi_1 n)$ samples drawn from $Q$ whereas $G_n$ will have all $n$ samples drawn from $Q$. Since $\pi_1 = 1 - \frac{1}{n-1}$, $\pi_1 n$ is always an integer for integer values of $n$. We denote samples that are drawn from $P$ as $p_i$ and samples drawn from $Q$ as $q_i$. Therefore, since $\{F_n\} = \{p_1, \ldots, p_{\pi_1 n}, q_{\pi_1 n + 1}, \ldots, q_n\}$, and $\{G_n\} = \{q_1, \ldots, q_n\}$, we decompose the empirical MMD distance of the mixture distribution as visualized in Fig. 5.

\[ D_{mmd}(\{F_n\}, \{G_n\}) = \frac{1}{n^2 - n}\left( \sum_{i=1}^{\pi_1 n} k(p_i, p_j) + \sum_{j=\pi_1 n + 1}^{n} k(p_i, q_j) \right) \]

\[+ \sum_{j=1}^{\pi_1 n} k(p_i, q_j) + \sum_{j=\pi_1 n + 1}^{n} k(q_i, q_j) - 2 \left( \sum_{i=1}^{\pi_1 n} \sum_{j=1, j \neq i}^{\pi_1 n} k(p_i, q_j) \right) \]

\[+ \left( \sum_{i=1}^{\pi_1 n} \sum_{j=1, j \neq i}^{\pi_1 n} k(p_i, q_j) \right). \hspace{1cm} (57)\]

We define $E[k(p_i, p_j)] = d_{pp}$, $E[k(p_i, q_j)] = E[k(q_i, p_j)] = d_{pq}$, and $E[k(q_i, q_j)] = d_{qq}$. The expectation of the estimator becomes:

$$E\left[ D_{mmd}(\{F_n\}, \{G_n\}) \right]$$

\[= \frac{1}{n^2 - n}\left( \left[ (\pi_1 n)^2 - \pi_1 n \right] d_{pp} + \left[ 2\pi_1 (1 - \pi_1) n^2 - 2(\pi_1 n^2 - \pi_1 n) \right] \right) \]

\[+\left(1 - \pi_1)^2 n^2 - (1 - \pi_1) n - 2(1 - \pi_1) n^2 - (1 - \pi_1) n \right) + (n^2 - 1) \]

\[= \frac{1}{(n^2 - n)}\left( d_{pp} + d_{qq} - 2d_{pq} \right). \hspace{1cm} (58)\]

The symmetric property of the MMD follows from the symmetry of the kernel as $E[k(p_i, q_j)] = E[k(q_i, p_j)]$.

From the definition of the WQT,

$$D_{wqt}(F_n, G_n) = \frac{n}{2} \int_0^1 \left( F_n(G_{n-1}(x)) - x \right)^2 dx. \hspace{1cm} (60)$$

The quantile-quantile map $F(G^{-1}(x))$ is a monotone function where $F(G^{-1}(0)) = 0$ and $F(G^{-1}(1)) = 1$. Therefore, the WQT is maximized in two cases: when $F(G^{-1}(x)) \equiv 0$ for all $x \in [0, 1)$ or $F(G^{-1}(x)) \equiv 1$ for all $x \in (0, 1]$. In each case,

$$W_{wqt} = \frac{n}{6}. \hspace{1cm} (61)$$

This occurs only when the 0-th quantile of $F$ maps to the 100-th quantile of $G$ or vice versa.

**References**

[1] A. Wald, *Sequential Analysis*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. J. Wiley & Sons, Incorporated.

[2] W. A. Shephard, *Economic control of quality of manufactured product*. Economic control of quality of manufactured product., Van Nostrand.

[3] M. Basseville and I. V. Nikiforov, “Detection of abrupt changes: Theory and application,” p. 469.

[4] F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, and Y. Amirat, “Joint segmentation of multivariate time series with hidden process regression for human activity recognition,” vol. 120, pp. 633–644.

[5] W.-H. Lee, J. Ortiz, B. Ko, and R. Lee, “Time series segmentation through automatic feature learning.”

[6] A. Willsky and H. Jones, “A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems,” vol. 21, no. 1, pp. 108–112.

[7] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, “Change-point detection in time-series data by relative density-ratio estimation,” vol. 43, pp. 72–83.

[8] J.-P. Qi, Q. Zhang, Y. Zhu, and J. Qi, “A novel method for fast change-point detection on simulated time series and electrocardiogram data,” vol. 9, no. 4, p. e93365.

[9] A. S. Dizaji and H. Soltanian-Zadeh, “A change-point analysis method for single-trial study of simultaneous EEG-MRI of auditory/visual oddball task,” p. 100487.

[10] S. Li, Y. Xie, H. Dai, and L. Song, “M-statistic for kernel change-point detection,” in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 3366–3374, Curran Associates, Inc.
Fig. 8. Decomposition of the expected value of the empirical MMD between $F_n$ and $G_n$ in (57). $d_{pp} = E[p_i,p_j]$, $d_{pq} = E[p_i,q_j]$, $d_{qq} = E[q_i,q_j]$ where $p_i \sim P$, and $q_i \sim Q$. 

[11] J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, “A review and comparison of change point detection techniques for climate data,” vol. 46, no. 6, pp. 900–915.
[12] D. M. Hawkins and Q. Deng, “A nonparametric change-point control chart,” vol. 42, no. 2, pp. 165–173.
[13] G. J. Ross and N. M. Adams, “Two nonparametric control charts for detecting arbitrary distribution changes,” vol. 44, no. 2, pp. 102–116.
[14] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, “On the empirical estimation of integral probability metrics,” vol. 6, no. 0, pp. 1550–1599.
[15] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” vol. 13, pp. 723–773.
[16] M. Sommerfeld and A. Munk, “Inference for empirical wasserstein distances on finite spaces,”
[17] C. Truong, L. Oudre, and N. Vayatis, “Selective review of offline change point detection methods,” vol. 167, p. 107299.
[18] W.-C. Chang, C.-L. Li, Y. Yang, and B. Pezos, “Kernel change-point detection with auxiliary deep generative models,”
[19] H. Ichino, K. Kaji, K. Sakurada, K. Hiroi, and N. Kawaguchi, “HASC-PAC2016: large scale human pedestrian activity corpus and its baseline recognition,” in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp ’16, pp. 705–714, ACM Press.
[20] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, “Learning and inferring motion patterns using parametric segmental switching linear dynamic systems,” vol. 77, no. 1, pp. 103–124.
[21] A. V. Oppenheim and G. C. Verghese, Signals, systems & inference. Pearson.
[22] A. Ramdas, N. G. Trillos, and M. Cuturi, “On wasserstein two-sample testing and related families of nonparametric tests,” vol. 19, p. 47.
[23] G. Peyr and M. Cuturi, “Computational optimal transport,”
[24] N. Bonneel, J. Rabin, G. Peyr, and H. Pfister, “Sliced and radon wasserstein barycenters of measures,” vol. 1, no. 51, pp. 22–45.
[25] F. J. Massey, “The kolmogorov-smirnov test for goodness of fit,” vol. 46, no. 253, pp. 68–78.
[26] N. Stewart, “Computing the kolmogorov-smirnov distribution when the underlying cdf is purely discrete, mixed or continuous,” p. 44.
[27] Y. Kawahara and M. Sugiyama, “Sequential change-point detection based on direct density-ratio estimation,” vol. 5, no. 2, pp. 114–127.
[28] A. W. v. d. Vaart, Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics, Cambridge University Press.
[29] D. Pollard, Convergence of stochastic processes. Springer series in statistics, Springer-Verlag.
[30] P. Billingsley, Convergence of probability measures. Wiley Series in Probability and Statistics. Probability and Statistics Section, John Wiley & Sons, 2nd ed. ed.
[31] A. W. v. d. Vaart and J. A. Wellner, Weak convergence and empirical processes. Springer series in statistics, Springer.
[32] E. Key, “Disks, shells, and integrals of inverse functions,” vol. 25, no. 2, pp. 136–138.
[33] L. Tolmatz, “On the distribution of the square integral of the brownian bridge,” vol. 30, no. 1, pp. 253–269.