A Another Solution to Tasks II and I

This section introduces an alternative method for solving tasks II and I. The proposed method is simpler and more intuitive than KCE and KCD, but does not have the same theoretical guarantees. Still, its empirical performance in the datasets we investigated is better than directly applying existing methods. We present it here as a candidate alternative for solving tasks I and II.

A.1 Another Solution to Task II

We start by fitting one function $\hat{f}$ using all the data $\{(x_t, y_t), t = 1, 2, \cdots, n\}$. For simplicity, we restrict our attention to Nadaraya-Watson estimator (3). When the null is true, all observations are i.i.d. samples from the same distribution. Under some mild conditions, we expect $\hat{f}$ to be close to the truth, and thus the residual $y_t - \hat{f}(x_t)$ distribution should be close to F_ϵ^0. However, when the alternative is true, one might expect that the fitted function $\hat{f}$ is a mix of f_0 and f_1, and thus the residual $\hat{\epsilon}_t = y_t - \hat{f}(x_t)$ will exhibit some relevant pattern. This intuition is demonstrated in the left two panels of Figure 3, where we plot the residuals $\hat{\epsilon}_t$ against t and observe that before change point ($\tau^* = 500$), $\hat{\epsilon}_t$'s roughly follow the same distribution, while after change point, $\hat{\epsilon}_t$'s follow a different distribution. This observation is confirmed by the following theorem; we denote F^Δ to be the probability distribution such that $f_1(X) - f_0(X) \sim F^\Delta$ with $X \sim F_X$.

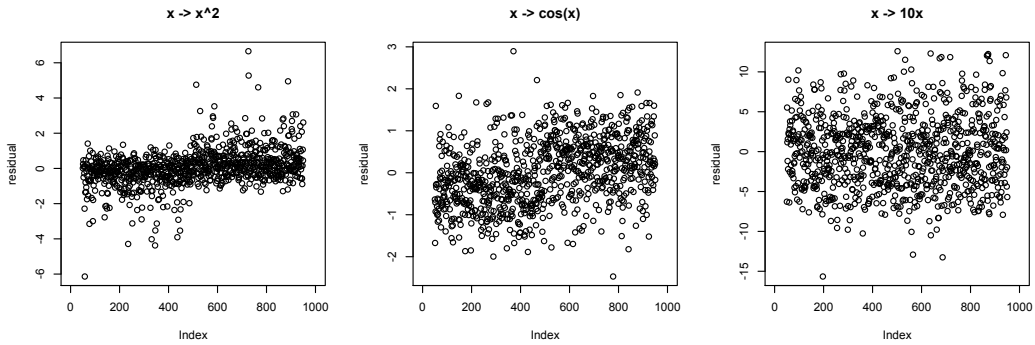

Figure 3: Plot of residuals $\hat{\epsilon}_t = y_t - \hat{f}(x_t)$ against t, where $\hat{f}$ is the Nadaraya-Watson estimator fitted using all the observations $\{(x_t, y_t), t = 1, 2, \cdots, n\}$. In all three panels we have $F_X = N(0, 1)$, $f_0(x) = x$, $n = 1000$, $\rho^* = 0.5$, $F_\epsilon^0 = F_\epsilon^1 = N(0, 1)$. In the left panel $f_1(x) = x^2$, in the middle $f_1(x) = \cos(x)$, and in the right $f_1(x) = 10x$.

Theorem A.1 (Pointwise asymptotic distribution of $\hat{\epsilon}_t$)**.** *Under the same assumptions as in Theorem 5.1, we have*

(1) Under the null, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho_1]$, as $n \to \infty$,

$$\hat{\epsilon}_t \xrightarrow{d} F_\epsilon^0.$$

(2) Under the alternative, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho^*]$, as $n \to \infty$,

$$\hat{\epsilon}_t \xrightarrow{d} (\rho^* - 1) Z + \epsilon, \quad \textit{where} \quad Z \sim F^\Delta, \quad \epsilon \sim F_\epsilon^0.$$

And for any $t = \lceil n\rho \rceil$ with $\rho \in (\rho^*, \rho_1]$, as $n \to \infty$,

$$\hat{\epsilon}_t \xrightarrow{d} \rho^* Z + \epsilon', \quad \textit{where} \quad Z \sim F^\Delta, \quad \epsilon' \sim F_\epsilon^1.$$

Proof. See Section D.3. □

Theorem A.1 shows that asymptotically, the residuals are distributed as $(\rho^* - 1) Z + \epsilon$ before change point, and $\rho^* Z + \epsilon'$ after change point. The implication is that if (unfortunately)

$$(\rho^* - 1) Z + \epsilon \sim \rho^* Z + \epsilon', \tag{15}$$

the distribution of residual $\hat{\epsilon}_t$'s will be identical across t, and we do not expect any pattern among the residual $\hat{\epsilon}_t$'s (see the right panel in Figure 3). In this case, it would be impossible to solve the change point problem merely based on the residuals. One example where Equation (15) holds is

$$F_\epsilon^0 = F_\epsilon^1, \quad \rho^* = 1/2, \quad Z \sim -Z, \quad \text{i.e., } F^\Delta \text{ is a symmetric distribution.}$$

When Equation (15) does not hold, roughly speaking, the univariate sequence $\{\hat{\epsilon}_t\}_{t=1}^T$ will exhibit an abrupt change point at τ^* in terms of the distribution of $\hat{\epsilon}_t$'s. This immediately suggests the potential applicability of existing literature on univariate, abrupt change point problems. Performance of the final estimator inevitably depends on the subsequent procedure that is used. Here we provide an analysis where a simple CUSUM procedure [35] is applied, i.e., the estimator used for detection is

$$D_\epsilon = \max_{\lceil n\rho_0 \rceil \le t \le \lceil n\rho_1 \rceil} \frac{t(n-t)}{n} T_\epsilon(t), \quad \text{where} \quad T_\epsilon(t) = \left(\frac{1}{t} \sum_{i=1}^{t} \hat{\epsilon}_i - \frac{1}{n-t} \sum_{i=t+1}^{n} \hat{\epsilon}_i \right)^2. \tag{16}$$

The estimator for change point is defined as

$$\hat{\tau} = \arg\max_{\lceil n\rho_0 \rceil \le t \le \lceil n\rho_1 \rceil} \frac{t(n-t)}{n} T_\epsilon(t), \tag{17}$$

where (ρ_0, ρ_1) satisfies $0 < \rho_0 \le \rho^* \le \rho_1 < 1$ and is set a priori to avoid arbitrarily small intervals near the boundary. Denote $n_0 = \lceil n\rho_0 \rceil, n_1 = \lceil n\rho_1 \rceil$. The complete procedure for solving Task II using this estimator is called KCE (Residual Kernel-based change point analysis for Conditional Expectations) and is summarized in Algorithm 3.

Algorithm 3 RKCE.

input: observations $\{(x_t, y_t), t = 1, 2, \cdots, n\}$, significance level α, parameters n_0, n_1.
output: estimated change point location $\hat{\tau}$. ▷ $\hat{\tau} = n$ implies no significant change point.
computation:
  fit function $\hat{f}$ using all observations.
  calculate residuals $\hat{\epsilon}_t = y_t - \hat{f}(x_t),\ t = 1, 2, \cdots, n$.
detection & localization:
  Calculate D_ϵ using Equation (16).
  if $D_\epsilon > c$ where c is the threshold (determined by permutations or bootstrap):
    return $\hat{\tau}$ using Equation (17).
  else:
    return $\hat{\tau} = n$.

We note that when using Algorithm 3, we are essentially using the method in [33] with Euclidean distances, and we are investigating changes in $\mathbb{E}[\hat{\epsilon}_t]$ only. With Theorem A.1, this implies that as long as

$$\mathbb{E}[f_0(X)] = \mathbb{E}[f_1(X)], \quad \text{i.e.,} \quad \mathbb{E}Y_1 = \mathbb{E}Y_2 = \cdots \mathbb{E}Y_n,$$

Algorithm 3 will fail. It is interesting to compare Algorithm 3 against two baseline methods:

- D_Y, which uses for detection

$$D_Y = \max_{n_0 \le t \le n_1} \frac{t(n-t)}{n} T_Y(t), \quad \text{where}$$
$$T_Y(t) = \left(\frac{1}{t} \sum_{i=1}^{t} y_i - \frac{1}{n-t} \sum_{i=t+1}^{n} y_i \right)^2,$$

  and for localization

$$\hat{\tau} = \arg\max_{n_0 \le t \le n_1} \frac{t(n-t)}{n} T_Y(t).$$

  Notice D_Y also looks at changes in $\mathbb{E}Y_t$ only, but without the extra burden of doing any function fitting. It is intriguing to understand when (and why) Algorithm 3 might be better than this baseline method. It turns out that Algorithm 3 and this baseline method have the same signal strength, yet differ in terms of noise level. As long as

$$\text{Cov}(f_0(X), f_1(X)) \ge 0,$$

Algorithm 3 filters out noise in raw y_t's, and thus will be more powerful and accurate. Intuitively, this is because when $f_0(X)$ and $f_1(X)$ are positively correlated, fitting a function $\hat{f}$ helps remove their "common" part and focus on their difference, which is kept in $\hat{\epsilon}_t$. When $f_0(X)$ and $f_1(X)$ are negatively correlated, depending on the correlation coefficient and change point location ρ^*, it might be either Algorithm 3 or this baseline method that work better.

- D_{XY}, which uses for detection

$$D_{XY} = \max_{n_0 \le t \le n_1} \frac{t(n-t)}{n} T_{XY}(t), \quad \text{where}$$
$$T_{XY}(t) = \left\| \frac{1}{t}\sum_{i=1}^{t} z_i - \frac{1}{n-t}\sum_{i=t+1}^{n} z_i \right\|_2^2,$$

with $z_t = (x_t', y_t)'$ and we assume $\mathcal{X} \subset \mathbb{R}^p$. For localization, we use

$$\hat{\tau} = \arg\max_{n_0 \le t \le n_1} \frac{t(n-t)}{n} T_{XY}(t).$$

The difference to D_Y is that D_{XY} treats (x_t, y_t) as a whole. So D_{XY} searches for changes in both $\mathbb{E}Y_t$ and $\mathbb{E}X_t$, and we expect the signal strength to be the same as D_ϵ. However, what differs (again) is the noise level. We expect D_{XY} to be more variable than D_ϵ, which implies that the localization error of D_{XY} might be larger than that of D_ϵ.

Empirical studies support our intuition for the comparison between D_ϵ, D_{XY}, D_Y. Moreover, in empirical studies, we find that the seemingly simple Algorithm 3 is surprisingly powerful and accurate for many settings. However, as discussed, there are cases where it fails. So in general we recommend using KCE instead of RKCE for solving task II.

A.2 Another Solution to Task I

Notice that the residual-based statistic $T_\epsilon(t)$ defined in (16) can be equivalently written as

$$T_\epsilon(t) = \left(\frac{1}{t}\sum_{i=1}^{t} (y_i - \hat{f}(x_i)) - \frac{1}{n-t}\sum_{i=t+1}^{n} (y_i - \hat{f}(x_i)) \right)^2.$$

Plugging the Nadaraya-Watson estimator (3) into the above expression, expanding the squared form and re-organizing it, we obtain

$$\begin{aligned} T_\epsilon(t) =& \frac{1}{t^2}\sum_{i,j=1}^{t} w(t,i,x_i)w(t,j,x_j)y_i y_j + \frac{1}{(n-t)^2}\sum_{i,j=t+1}^{n} w(t,i,x_i)w(t,j,x_j)y_i y_j \\ & - \frac{2}{t(n-t)}\sum_{i=1}^{t}\sum_{j=t+1}^{n} w(t,i,x_i)w(t,j,x_j)y_i y_j \end{aligned} \tag{18}$$

where we denote for simplicity $k_X\left(h_X^{-1} d(x_i, x_j)\right) = k_X(i,j)$, and

$$w(t,i,x_i) = \begin{cases} \frac{1/(n-t)\sum_{j=t+1}^{n} k_X(i,j)}{(1/n)\sum_{j=1}^{n} k_X(i,j)}, & \text{if } i \le t, \\ \frac{(1/t)\sum_{j=1}^{t} k_X(i,j)}{(1/n)\sum_{j=1}^{n} k_X(i,j)}, & \text{if } i > t. \end{cases} \tag{19}$$

By replacing the inner product $y_i y_j$ by kernels $k_Y(y_i, y_j)$, it is easy to generalize Equation (18) to

$$\begin{aligned} T_\epsilon(t) =& \frac{1}{t^2}\sum_{i,j=1}^{t} w(t,i,x_i)w(t,j,x_j)k_Y(y_i,y_j) + \frac{1}{(n-t)^2}\sum_{i,j=t+1}^{n} w(t,i,x_i)w(t,j,x_j)k_Y(y_i,y_j) \\ & - \frac{2}{t(n-t)}\sum_{i=1}^{t}\sum_{j=t+1}^{n} w(t,i,x_i)w(t,j,x_j)k_Y(y_i,y_j), \end{aligned} \tag{20}$$

where $w(t, i, x_i)$ is defined in Equation (19). Using (20), we can solve task I and the complete procedure is summarized in Algorithm 4. By choosing k_Y as universal kernels, analogous to Task II which is essentially looking for changes over $\mathbb{E}Y$, under appropriate conditions, Algorithm 3 will be analyzing changes in the whole distribution of Y. For instance, if choosing Gaussian kernel $k_Y(y, y') = \exp\{-(y - y')^2/h_Y^2\}$ for $\mathcal{Y} = \mathbb{R}$ and some fixed bandwidth $h > 0$, we are looking at changes in any of $\{\mathbb{E}Y^l, l = 1, 2, \cdots\}$, which, under appropriate conditions, is equivalent to changes in distribution of Y. Algorithm (4) is named RKCD (Residual Kernel-based change point analysis for Conditional Distributions).

Algorithm 4 RKCD.

input: observations $\{(x_t, y_t), t = 1, 2, \cdots, n\}$, significance level α, parameters n_0, n_1.
output: estimated change point location $\hat{\tau}$. ▷ $\hat{\tau} = n$ implies no significant change point.
pre-compute:
1. $K_X = [k_X(h_X^{-1} d(x_i, x_j))]_{i,j=1}^n \in \mathbb{R}^{n\times n}$, $K_Y = [k_Y(y_i, y_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n}$.
2. $A \in \mathbb{R}^{n\times n}$ where $A_{ij} = \sum_{l=1}^i \sum_{m=1}^j [K_X]_{lm}$.
for $t = n_0, n_0 + 1, \cdots, n_1$ **do**
calculate $w \in \mathbb{R}^n$ with

$$w_i = \begin{cases} \frac{n}{n-t} \frac{B_{in} - B_{it}}{B_{in}}, & \text{if } i \le t, \\ \frac{n}{t} \frac{B_{it}}{B_{in}}, & \text{if } i > t. \end{cases}$$

calculate $Q \in \mathbb{R}^{n\times n}$ with $Q_{ij} = w_i w_j [K_Y]_{ij}$.
calculate $T_\epsilon = \frac{1}{t^2} \sum_{i,j=1}^t Q_{ij} + \frac{1}{(n-t)^2} \sum_{i,j=t+1}^n Q_{ij} - \frac{2}{t(n-t)} \sum_{i=1}^t \sum_{j=t+1}^n Q_{ij}$.
end for
detection: obtain p-value for $\max_{n_0 \le t \le n_1} [t(n-t)/n] T_\epsilon$ using permutations or bootstrap.
localization: if p-value $< \alpha$, estimate $\hat{\tau} = \arg\max_{n_0 \le t \le n_1} [t(n-t)/n] T_\epsilon$; else, estimate $\hat{\tau} = n$.

B More Related Work

This section discusses the relationship between this work and the literature on kernel embeddings of conditional distributions [36, 41]. Recall that Remark 5.4 established the connection between Δ defined in (13) and the maximal conditional mean discrepancy (MCMD) in [36]. A by-product of this work is that, similarly to the derivation of (12), we can obtain a new estimate for the MCMD.

Denote $F^0_{Y|X=x}, F^1_{Y|X=x}$ two probability distributions whose MCMD we aim to estimate. Let $x_i, y_i, i = 1, 2, \cdots, m$ be a sample drawn from $F^0_{Y|X=x}$ and $x'_i, y'_i, i = 1, 2, \cdots, n$ from $F^1_{Y|X=x}$. Then we can estimate the squared MCMD between $F^0_{Y|X=x}, F^1_{Y|X=x}$ as

$$\sum_{i,j=1}^m w_0(i) w_0(j) k_Y(y_i, y_j) + \sum_{i,j=1}^n w_1(i) w_1(j) k_Y(y'_i, y'_j) - 2 \sum_{i=1}^m \sum_{j=1}^n w_0(i) w_1(j) k_Y(y_i, y'_j), \tag{21}$$

where

$$\begin{cases} w_0(i) = \frac{k_X\left(h_X^{-1} d(x_i, x)\right)}{\sum_{r=1}^m k_X\left(h_X^{-1} d(x_r, x)\right)}, \\ w_1(i) = \frac{k_X\left(h_X^{-1} d(x'_i, x)\right)}{\sum_{r=1}^n k_X\left(h_X^{-1} d(x'_r, x)\right)}. \end{cases} \tag{22}$$

Comparing (21) with the empirical estimator proposed in [36], we notice that (21) has the following advantages: (i) it is computationally cheaper, with a $O((m + n)^2)$ time complexity rather than the $O((m + n)^3)$ for the estimator in [36]; (ii) it directly estimates MCMD and does not involve any surrogate loss as in [36], which could bring additional errors.

The conditional maximum mean discrepancy (CMMD) ([41]) is another measure for the discrepancy between two conditional distributions. However, as discussed in [36], CMMD often either does not exist or is not an exact measure of discrepancy at the population level.

We also note that MCMD and CMMD can both be viewed as generalizations of the MMD [17] which is designed for testing the difference of two unconditional distributions.

C Additional Theoretical Results

This section presents theory related to Task II, i.e, $\mathcal{Y} \subset \mathbb{R}, \mathcal{X} \subset \mathbb{R}^p$, $\mathbb{E}[y_t \mid x_t]$ changes, and $\tilde{\Delta}_t$ is defined in Equation (6). First, let us discuss some notations and assumptions.

Assumption 6. *The set $\mathcal{X}$ is a bounded subset of $\mathbb{R}^q$.*

Assumption 7. *The kernel $k_X\left(h_X^{-1}d(x,x')\right) = K\left(\frac{x-x'}{h_X}\right)$ where $K : \mathbb{R}^q \to [0,\infty)$ is a multivariate uth-order kernel function with*

$$|K(x)| \le c_2 < \infty, \quad \int_{\mathbb{R}^q} |K(x)|\,dx < \infty, \quad \int_{\mathbb{R}^q} |x|^u\,|K(x)|\,dx < \infty.$$

Assumption 8. *For some $s > 2$, $\mathbb{E}[|Y|^s] < \infty$. Covariate X has marginal density $p(x)$ such that*

$$0 < c_0 \le \inf_{x\in\mathcal{X}} p(x) \le \sup_{x\in\mathcal{X}} p(x) \le c_1 < \infty,$$

and

$$\sup_{x\in\mathcal{X}} \mathbb{E}\left[|Y|^s \mid X = x\right] p(x) \le c_2 < \infty.$$

Assumption 9. *The second derivatives of $p(x)$ and $f_0(x)p(x)$, $f_1(x)p(x)$ are uniformly continuous and bounded. The uth derivative of $p(x)$ is uniformly continuous.*

Assumption 10. *The bandwidth satisfies $\frac{\log n}{nh_X^q} = o(1)$ and $h_X = o(1)$.*

Denote $\xrightarrow{p}$ as convergence in probability. We have the following theorem:

Theorem C.1. *Suppose Assumptions 6, 7, 8, 9, 10 hold. Then the following hold:*

(1) Under the null, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho_1]$, as $n \to \infty$,

$$\tilde{\Delta}_t \xrightarrow{p} 0.$$

(2) Under the alternative, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho^*) \cap (\rho^*, \rho_1]$,

$$\tilde{\Delta}_t \xrightarrow{p} \delta(\rho)\mathbb{E}_{X\sim F_X}[f_0(X) - f_1(X)]^2,$$

where $\delta(\cdot)$ defined in Equation (14).

Proof. See Appendix D.1. □

Remark C.1. *Theorem C.1 shows that the sequence $\{\tilde{\Delta}_t, t = 1, 2, \cdots, n\}$ satisfies the properties shown in Figure 1 (i.e., flat across all t's under the null, and large around $t = \tau^*$ under the alternative). Morevoer, it is interesting to compare Theorem C.1 against Theorem A.1. One difference is that, if using Algorithm 3, performance of RKCE depends crucially on*

$$\mathbb{E}[f_0(X) - f_1(X)].$$

This quantity can be zero, even if a change actually occurs. And in this case, RKCE will fail. In contrast, performance of KCE depends on

$$\mathbb{E}[(f_0(X) - f_1(X))^2],$$

which should always *be nonzero as long as f_0, f_1 are continuous, and differ in at least one point in the support of X.*

Proposition 2. *When the null is true, let $c_1(h_X) > 0$ be some constant that depends on the bandwidth h_X of $\hat{f}$. Under appropriate choice of h_X, there exists positive constant $c_2(x)$ depending on x such that for any fixed $x > 0$, as $n \to \infty$,*

$$\sqrt{n}c_1(h_X)\left(\hat{f}(x) - f_0(x)\right) \xrightarrow{d} N(0, c_2(x)).$$

Remark C.2. *Proposition 2 has been demonstrated under some mild assumptions in nonparametric literature. See, for example [45] for the case where $\mathcal{X} \subset \mathbb{R}$, [5] for $\mathcal{X} \subset \mathbb{R}^p$, and [3] for general $\mathcal{X}$'s. The assumptions are standard so we will not include them here for brevity. We give an example where $\mathcal{X} \subset \mathbb{R}^q$. Then under conditions in Theorem 6.2.1 of [5], Proposition 2 holds with $c_1(h_X) = \sqrt{h_X^q}$, and*

$$c_2(x) = \frac{[R(k_X)]^q Var(Y \mid X)}{p(x)}, \quad with \quad R(k_X) = \int [k_X(u)]^2 du.$$

D Technical Proofs

This section contains the proof to Theorem C.1, Theorem 5.1, and Theorem A.1.

Notations. Denote $\xrightarrow{p}$ as convergence in probability. For a set of random variables X_n and a corresponding set of constants a_n, denote $X_n = o_p(a_n)$ when X_n/a_n converges to zero in probability. Recall the notion of almost complete convergence introduced in [14]. Following their notation, we write $X_n = o_{\text{a.co.}}(a_n)$ if $\forall \epsilon > 0$, $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n/a_n| > \epsilon) < \infty$, and write $X_n = O_{\text{a.co.}}(a_n)$ if $\exists \epsilon > 0$ such that $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n/a_n| > \epsilon_0) < \infty$. We also write $X_n \xrightarrow{a.co.} 0$ when $X_n = o_{\text{a.co.}}(1)$.

D.1 Proof to Theorem C.1

Proof. Note that the null case can be viewed as a special case of the alternative with $\Delta = 0$. Thus, we will focus on proof under the alternative. We will only present the proof for $\rho < \rho^*$. The proof for $\rho > \rho^*$ is similar.

Recall that for any fixed $\rho \in [\rho_0, \rho_1]$ and $t = \lceil T\rho \rceil$, $\tilde{\Delta}_t$ is defined as

$$\tilde{\Delta}_t = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}_-(t, x_i) - \hat{f}_+(t, x_i)]^2.$$

Denote

$$g_t(x_i) = \hat{f}_-(t, x_i) - \hat{f}_+(t, x_i) - \frac{1-\rho^*}{1-\rho} f^*(x_i),$$

where $f^* = f_0 - f_1$. Then we have

$$\begin{aligned}\tilde{\Delta}_t &= \frac{1}{n}\sum_{i=1}^{n}\left[g_t(x_i) + \frac{1-\rho^*}{1-\rho} f^*(x_i)\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[g_t(x_i)]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-\rho^*}{1-\rho} f^*(x_i)\right]^2 + 2\frac{1-\rho^*}{1-\rho}\frac{1}{n}\sum_{i=1}^{n} g_t(x_i) f^*(x_i). \end{aligned} \tag{23}$$

Notice that from Lemma D.1,

$$\frac{1}{n}\sum_{i=1}^{n}[g_t(x_i)]^2 \le \|g_t\|_\infty^2 \xrightarrow{p} 0. \tag{24}$$

From the Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-\rho^*}{1-\rho} f^*(x_i)\right]^2 \xrightarrow{p} \left(\frac{1-\rho^*}{1-\rho}\right)^2 \Delta. \tag{25}$$

And from the uniform boundedness of f^* (from Assumption 6, 9) and (24), we have

$$\frac{1}{n}\sum_{i=1}^{n} g_t(x_i) f^*(x_i) \le \|g_t\|_\infty \|f^*\|_\infty \xrightarrow{p} 0. \tag{26}$$

Plugging (24), (25), (26) into (23), we have

$$\tilde{\Delta}_t \xrightarrow{p} \left(\frac{1-\rho^*}{1-\rho}\right)^2 \Delta.$$

□

Lemma D.1. *Suppose Assumptions 6, 7, 8, 9, 10 hold. Then under the alternative, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho^*)$, we have*

$$\left\|\hat{f}_-(\cdot, t) - f_0(\cdot)\right\|_\infty \xrightarrow{p} 0, \quad \left\|\hat{f}_+(\cdot, t) - \frac{1-\rho}{1-\rho^*} f_0(\cdot) - \frac{\rho-\rho^*}{1-\rho^*} f_1(\cdot)\right\|_\infty \xrightarrow{p} 0.$$

For any $t = \lceil n\rho \rceil$ with $\rho \in (\rho^*, \rho_1]$, we have

$$\left\|\hat{f}_-(\cdot, t) - \frac{\rho^*}{\rho} f_0(\cdot) - \left(1 - \frac{\rho^*}{\rho}\right) f_1(\cdot)\right\|_\infty \xrightarrow{p} 0, \quad \left\|\hat{f}_+(\cdot, t) - f_1(\cdot)\right\|_\infty \xrightarrow{p} 0.$$

Proof. For brevity we only prove the uniform consistency for $\hat{f}_-(t,\cdot)$ when $t = \lceil n\rho \rceil$ with $\rho \in (\rho^*, \rho_1]$. The other results can be proved similarly. Denote

$$f^*(\cdot) = \frac{\rho^*}{\rho} f_0(\cdot) - \left(1 - \frac{\rho^*}{\rho}\right) f_1(\cdot).$$

With a slight abuse of notation, we write $k_X(x, x') = k_X\left(h_X^{-1} d(x, x')\right)$. Notice that

$$\begin{aligned}
&\left\| \hat{f}_-(t,\cdot) - f^* \right\|_\infty \\
=& \left\| \frac{\sum_{i=1}^t k_X(\cdot, x_i) y_i}{\sum_{i=1}^t k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} f_0(\cdot) - \left(1 - \frac{\rho^*}{\rho}\right) f_1(\cdot) \right\|_\infty \\
=& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} + \frac{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \frac{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i) y_i}{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i)} \right. \\
&\qquad \left. - \frac{\rho^*}{\rho} f_0 - \left(1 - \frac{\rho^*}{\rho}\right) f_1 \right\|_\infty \\
\leq& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} f_0 \right\|_\infty \\
&\quad + \left\| \frac{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \frac{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i) y_i}{\sum_{i=\tau^*+1}^t k_X(\cdot, x_i)} - \left(1 - \frac{\rho^*}{\rho}\right) f_1 \right\|_\infty .
\end{aligned} \tag{27}$$

We will show that the first term in the above inequality converges in probability to 0. The second term can be proved similarly. Notice that

$$\begin{aligned}
&\left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} f_0 \right\|_\infty \\
\leq& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} \left(\frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - f_0 \right) \right\|_\infty + \left\| \left(\frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} \right) f_0 \right\|_\infty \\
\overset{(a)}{\leq}& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - f_0 \right\|_\infty + C \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^t k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} \right\|_\infty \\
\leq& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - f_0 \right\|_\infty + C \left\| \frac{\tau^*}{t} \frac{\frac{1}{\tau^* h^q} \sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\frac{1}{th^q} \sum_{i=1}^t k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} \right\|_\infty \\
\leq& \left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - f_0 \right\|_\infty + C \left\| \frac{\frac{1}{\tau^* h^q} \sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\frac{1}{th^q} \sum_{i=1}^t k_X(\cdot, x_i)} - 1 \right\|_\infty + O\left(n^{-1}\right),
\end{aligned} \tag{28}$$

where (a) follows from the non-negativity of k_X (Assumption 7) and the boundedness of f_0 (Assumption 6 and 9). Thus, if we could prove the two terms in (28) both converge in probability to 0, the proof is complete.

Let us first show that

$$\left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - f_0 \right\|_\infty \xrightarrow{p} 0. \tag{29}$$

This is a direct consequene of Theorem 8 in [18] and Assumption 9, 8, 10.

Second let us show

$$\left\| \frac{\frac{1}{\tau^* h^q} \sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\frac{1}{th^q} \sum_{i=1}^t k_X(\cdot, x_i)} - 1 \right\|_\infty \xrightarrow{p} 0. \tag{30}$$

Notice from Assumptions 7, 8, 10 and Theorem 6 in [18], we have

$$\left\| \frac{1}{\tau^* h^q} \sum_{i=1}^{\tau^*} k_X(\cdot, x_i) - p(x) \right\|_\infty = o_p(1),$$
$$\left\| \frac{1}{t h^q} \sum_{i=1}^{t} k_X(\cdot, x_i) - p(x) \right\|_\infty = o_p(1).$$

These two equations directly imply Equation (30).

Plugging (29) and (30) into (28), we have

$$\left\| \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)}{\sum_{i=1}^{t} k_X(\cdot, x_i)} \frac{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i) y_i}{\sum_{i=1}^{\tau^*} k_X(\cdot, x_i)} - \frac{\rho^*}{\rho} f_0 \right\|_\infty = o_p(1).$$

And similarly, one can show the second term in (27) is also $o_p(1)$. This concludes

$$\left\| \hat{f}_-(t, \cdot) - f^* \right\|_\infty = o_p(1).$$

□

D.2 Proof of Theorem 5.1

In this subsection, for notational brevity, we sometimes write $\|\cdot\|_{\mathcal{H}} = \|\cdot\|$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle$.

Proof. Conclusion (1) is a direct consequence of Lemma D.2. Now let us prove conclusion (2).

$$\begin{aligned}
&\tilde{\Delta}_t - \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{\rho^*}{\rho} \left[f_0(X_i) - f_1(X_i) \right] \right\|^2 \\
=& \frac{1}{n} \sum_{i=1}^{n} \Bigg\langle \hat{f}_-(X_i, t) - \hat{f}_+(X_i, t) - \frac{\rho^*}{\rho} \left(f_0(X_i) - f_1(X_i) \right), \\
&\qquad \hat{f}_-(X_i, t) - \hat{f}_+(X_i, t) + \frac{\rho^*}{\rho} \left(f_0(X_i) - f_1(X_i) \right) \Bigg\rangle \\
\overset{(a)}{\leq}& \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{f}_-(X_i, t) - \hat{f}_+(X_i, t) - \frac{\rho^*}{\rho} \left[f_0(X_i) - f_1(X_i) \right] \right\| \\
&\qquad \times \left\| \hat{f}_-(X_i, t) - \hat{f}_+(X_i, t) + \frac{\rho^*}{\rho} \left[f_0(X_i) - f_1(X_i) \right] \right\| \\
\leq& \sup_{x \in \mathcal{X}} \left\| \hat{f}_-(x, t) - \hat{f}_+(x, t) - \frac{\rho^*}{\rho} \left[f_0(x) - f_1(x) \right] \right\| \\
&\quad \times \sup_{x \in \mathcal{X}} \left\| \hat{f}_-(x, t) - \hat{f}_+(x, t) + \frac{\rho^*}{\rho} \left[f_0(x) - f_1(x) \right] \right\| \\
\overset{(b)}{\leq}& o_{\text{a.co.}}(1) \times \left[\sup_{x \in \mathcal{X}} \left\| \hat{f}_-(x, t) - \hat{f}_+(x, t) - \frac{\rho^*}{\rho} \left[f_0(x) - f_1(x) \right] \right\| + \sup_{x \in \mathcal{X}} \left\| 2 \frac{\rho^*}{\rho} \left[f_0(x) - f_1(x) \right] \right\| \right] \\
\overset{(c)}{=}& o_{\text{a.co.}}(1) \overset{(d)}{=} o_{\text{a.s.}}(1),
\end{aligned} \tag{31}$$

where (a) follows from Cauchy-Schwarz Inequality, (b) follows from Lemma D.2, (c) follows from Lemma D.2 and the boundedness of f_0, f_1 (Assumption 4), (d) follows from Proposition A.2 in [14]. Also from the strong law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^{n} \left\| \left(f_0(X_i) - f_1(X_i) \right) \right\|^2 - \mathbb{E} \left\| \left(f_0(X_i) - f_1(X_i) \right) \right\|^2 = 0, \quad \text{a.s.}. \tag{32}$$

Combining Equation (31) and Equation (32), we get

$$\tilde{\Delta}_t - \left(\frac{\rho^*}{\rho}\right)^2 \mathbb{E}\left\|\left(f_0(X) - f_1(X)\right)\right\|^2 = o_{\text{a.s.}}(1).$$

Similarly, we can show that for any $t = \lceil n\rho \rceil$ with $\rho \in (\rho^*, \rho_1]$,

$$\tilde{\Delta}_t - \left(\frac{1-\rho^*}{1-\rho}\right)^2 \mathbb{E}\left\|\left(f_0(X) - f_1(X)\right)\right\|^2 = o_{\text{a.s.}}(1).$$

This completes proof of the desired conclusion. □

Lemma D.2. *Suppose assumptions 1, 2, 3, 4, 5 hold.*
(1) Under the null, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho_1]$,

$$\sup_{x\in\mathcal{X}} \left\|\hat{f}_-(x,t) - \hat{f}_+(x,t)\right\|_{\mathcal{H}} = O_{a.co.}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}} + \sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right).$$

(2) Under the alternative, for any $t = \lceil n\rho \rceil$ with $\rho \in [\rho_0, \rho_1]$,

$$\sup_{x\in\mathcal{X}} \left\|\hat{f}_-(x,t) - \hat{f}_+(x,t) - \delta^{1/2}(\rho)\left(f_0(x) - f_1(x)\right)\right\|_{\mathcal{H}} = o_{a.co.}(1),$$

with $\delta(\cdot)$ defined in Equation (14).

Proof. Without loss of generality, Denote

$$\hat{p}_-(x,t) = \frac{\sum_{i=1}^t k_X\left(h^{-1}d(x,X_i)\right)}{t\mathbb{E}\left[k_X\left(h^{-1}d(x,X_1)\right)\right]}, \quad \hat{g}_-(x,t) = \frac{\sum_{i=1}^t k_X\left(h^{-1}d(x,X_i)\right)Y_i}{t\mathbb{E}\left[k_X\left(h^{-1}d(x,X_1)\right)\right]},$$
$$\hat{p}_+(x,t) = \frac{\sum_{i=t+1}^n k_X\left(h^{-1}d(x,X_i)\right)}{(n-t)\mathbb{E}\left[k_X\left(h^{-1}d(x,X_1)\right)\right]}, \quad \hat{g}_+(x,t) = \frac{\sum_{i=t+1}^n k_X\left(h^{-1}d(x,X_i)\right)Y_i}{(n-t)E\left[k_X\left(h^{-1}d(x,X_1)\right)\right]}.$$

Following [13], let us consider the following decomposition

$$\begin{aligned}
&\hat{f}_-(x,t) - \hat{f}_+(x,t) \\
=&\left[\frac{\hat{g}_-(x,t) - \mathbb{E}\hat{g}_-(x,t)}{\hat{p}_-(x,t)} + \frac{\mathbb{E}\hat{g}_-(x,t) - f(x)}{\hat{p}_-(x,t)} + \frac{\left(1-\hat{p}_-(x,t)\right)f(x)}{\hat{p}_-(x,t)}\right] \\
&-\left[\frac{\hat{g}_+(x,t) - \mathbb{E}\hat{g}_+(x,t)}{\hat{p}_+(x,t)} + \frac{\mathbb{E}\hat{g}_+(x,t) - f(x)}{\hat{p}_+(x,t)} + \frac{\left(1-\hat{p}_+(x,t)\right)f(x)}{\hat{p}_+(x,t)}\right].
\end{aligned}$$

Under the null, this becomes

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\left\|\hat{f}_-(x,t)-\hat{f}_+(x,t)\right\| \\
=&\sup_{x\in\mathcal{X}}\left\|\left[\frac{\hat{g}_-(x,t)-\mathbb{E}\hat{g}_-(x,t)}{\hat{p}_-(x,t)}-\frac{\hat{g}_+(x,t)-\mathbb{E}\hat{g}_+(x,t)}{\hat{p}_+(x,t)}\right]\right. \\
&+\left(\mathbb{E}\hat{g}_-(x,t)-f(x)\right)\left(\frac{1-\hat{p}_-(x,t)}{\hat{p}_-(x,t)}-\frac{1-\hat{p}_-(x,t)}{\hat{p}_+(x,t)}\right) \\
&\left.+f(x)\left(\frac{1-\hat{p}_-(x,t)}{\hat{p}_-(x,t)}-\frac{1-\hat{p}_-(x,t)}{\hat{p}_-(x,t)}\right)\right\| \\
\overset{(a)}{=}&O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}}\right)+O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right) \\
&+O(h^b)O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}}\right)+O(h^b)O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right) \\
&+O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}}\right)+O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right) \\
=&O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}}\right)+O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right),
\end{aligned}
$$

where (a) follows from Lemma 3.2 and 3.3 in [13], and the following results stated in [12]:

$$
\sup_{x\in\mathcal{X}}|\hat{p}_-(x,t)-1|=O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log t}{t}\right)}{tm(h)}}\right),\quad\sum_{n=1}^{\infty}\mathbb{P}\left(\inf_{x\in\mathcal{X}}\hat{p}_-(x,t)<\frac{1}{2}\right)<\infty, \tag{33}
$$

$$
\sup_{x\in\mathcal{X}}|\hat{p}_+(x,t)-1|=O_{\text{a.co.}}\left(\sqrt{\frac{\psi_{\mathcal{X}}\left(\frac{\log(n-t)}{n-t}\right)}{(n-t)m(h)}}\right),\quad\sum_{n=1}^{\infty}\mathbb{P}\left(\inf_{x\in\mathcal{X}}\hat{p}_+(x,t)<\frac{1}{2}\right)<\infty. \tag{34}
$$

Thus, conclusion (1) holds.

Under the alternative, we prove for any $t=\lceil n\rho\rceil$ with $\rho\in(\rho^*,\rho_1]$, the following holds

$$
\sup_{x\in\mathcal{X}}\left\|\hat{f}_-(x,t)-w_0f_0(x)-(1-w_0)f_1(x)\right\|=o_{\text{a.co.}}(1), \tag{35}
$$

$$
\sup_{x\in\mathcal{X}}\left\|\hat{f}_+(x,t)-f_1(x)\right\|=o_{\text{a.co.}}(1), \tag{36}
$$

where $w_0=\frac{\rho^*}{\rho}$. Denote

$$
w=\frac{\sum_{i\le\tau^*}k_X\left(h^{-1}d(x,X_i)\right)}{\sum_{i\le t}k_X\left(h^{-1}d(x,X_i)\right)},
$$

and notice that from Equation (33) and Equation (34), we have

$$
w\xrightarrow{\text{a.co.}}w_0.
$$

First let us prove Equation (35), notice that from conclusion (1),

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left\|\hat{f}_-(x,t)-w_0f_0(x)-(1-w_0)f_1(x)\right\| \\
=&\sup_{x\in\mathcal{X}}\left\|\frac{\sum_{i=1}^{t}k_X\left(h^{-1}d(x,X_i)\right)k_Y(Y_i,\cdot)}{\sum_{i=1}^{t}k_X\left(h^{-1}d(x,X_i)\right)}-wf_0(x)-(1-w)f_1(x)\right. \\
&\qquad\left.+(w-w_0)f_0(x)+(w_0-w)f_1(x)\right\| \\
=&\sup_{x\in\mathcal{X}}\left\|w\left(\frac{\sum_{i=1}^{\tau^*}k_X\left(h^{-1}d(x,X_i)\right)k_Y(Y_i,\cdot)}{\sum_{i=1}^{\tau^*}k_X\left(h^{-1}d(x,X_i)\right)}-f_0(x)\right)\right. \\
&\qquad+(1-w)\left(\frac{\sum_{i=\tau^*+1}^{t}k_X\left(h^{-1}d(x,X_i)\right)k_Y(Y_i,\cdot)}{\sum_{i=\tau^*+1}^{t}k_X\left(h^{-1}d(x,X_i)\right)}-f_1(x)\right) \\
&\qquad\left.+(w-w_0)\left(f_0(x)-f_1(x)\right)\right\| \\
=&o_{\text{a.co.}}(1),
\end{aligned}$$

and similarly,

$$\sup_{x\in\mathcal{X}}\left\|\hat{f}_+(x,t)-f_1(x)\right\|=o_{\text{a.co.}}(1).$$

Thus, we have for any $t=\lceil n\rho\rceil$ with $\rho\in(\rho^*,\rho_1]$,

$$\sup_{x\in\mathcal{X}}\left\|\hat{f}_-(x,t)-\hat{f}_+(x,t)-\frac{\rho^*}{\rho}\left(f_0(x)-f_1(x)\right)\right\|=o_{\text{a.co.}}(1).$$

And similarly, we can prove for any $t=\lceil n\rho\rceil$ with $\rho\in[\rho_0,\rho^*)$,

$$\sup_{x\in\mathcal{X}}\left\|\hat{f}_-(x,t)-\hat{f}_+(x,t)-\frac{1-\rho^*}{1-\rho}\left(f_0(x)-f_1(x)\right)\right\|=o_{\text{a.co.}}(1).$$

□

D.3 Proof to Theorem A.1

Proof. Notice that (1) is a special case of (2). Thus, we will only show (2).

Similar to the proof of Equation (35), and from Proposition A.4 in [14], one can show that

$$\sup_{x\in\mathcal{X}}\left\|\hat{f}(x)-\rho^*f_0(x)-(1-\rho^*)f_1(x)\right\|_{\mathcal{H}}\xrightarrow{p}0.$$

Then conclusion (2) follows directly. □

D.4 Proof to Proposition 1

Proof. Suppose $F^0_{Y|X=x_0}\neq F^1_{Y|X=x_0}$ where $x_0\in\text{supp}(F_X)$. Since k_Y is a characteristic kernel, we have $f_0(x_0)\neq f_1(x_0)$ and thus, $\|f_0(x_0)-f_1(x_0)\|_{\mathcal{H}}=c>0$. Since $l:x\mapsto\|f_0(x)-f_1(x)\|_{\mathcal{H}}$ is a continuous function (from Assumption 2), there must exist an open neighborhood N_{x_0} of x_0 such that for any $x\in N_{x_0}$,

$$l(x)=\|f_0(x)-f_1(x)\|_{\mathcal{H}}>c/2.$$

Notice that

$$\begin{aligned}
\Delta=\mathbb{E}_{X\sim F_X}\|f_0(X)-f_1(X)\|^2_{\mathcal{H}}&=\int_{x\in\mathcal{X}}\|f_0(x)-f_1(x)\|^2_{\mathcal{H}}\,dF_X(x)\\
&>\int_{x\in N_{x_0}}\|f_0(x)-f_1(x)\|^2_{\mathcal{H}}\,dF_X(x)>(c/2)\,F_X\left(N_{x_0}\right)>0.
\end{aligned}$$

□

E Additional Experimental Results

This section contains additional results for power comparison in simulated datasets. As the fixed design method [31] does not perform well for our random design setting (as shown in Table 1a), it is omitted for this power comparison. Table 3a shows the results in experiment A, Table 3b experiment B, and Table 3c experiment C. Note that we make the alternatives more challenging to make power comparison informative. D_Y has the lowest power across all methods. The difference between the power of D_{XY} and KCD (or KCE) has been reduced, when compared with the localization results. But still, we find D_{XY} to be generally slightly worse than KCD (or KCE).

F Extension to multiple change points setting

This section discusses the generalization of the proposed method to multiple change points setting. We use a binary segmentation method, as outlined in Algorithm 5. Using this algorithm with $n_{\min} = 50, \alpha = 0.1, n'_0 = n'_1 = 5$, we sequentially find 3 change points: 2016/6/23, 2016/4/20, and 2016/2/26. The first change point corresponds to the date of Brexit vote, and the third corresponds to the announcement of Brexit vote. The second change point, however, has no clear association with any events in UK. Since we treat NIKKEI and NYSE as our covariates, our hypothesis is that this might have something to do with changes in NYSE around that date (2016/4/20), which marked the commanding victory of Donald Trump and Hillary Clinton in the New York primaries.

Algorithm 5 Generalization of KCE (or KCD) to multiple change points setting

input: observations $\{(x_t, y_t)\}_{t=1}^n$, significance level α, parameters n'_0, n'_1, minimum length $n_{\min}$.
output: Set of detected change points $\hat{\mathcal{D}} = \texttt{BS}(1, n)$. ($\hat{\tau} = n$ implies no significant change point)
Function: $\texttt{BS}(l, r)$
For the subsequence $\{x_l, \cdots, x_r\}$, use KCE or KCE (with significance level $\alpha, n_0 = n'_0$, and $n_1 = r - l - n'_1 + 1$) to find the new change point k.
If p-value of k is significant, and $k - l, r - k \geq n_{\min}$,
Update $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{k\}$.
Call $\texttt{BS}(l, k)$.
Call $\texttt{BS}(k + 1, r)$.
Return $\hat{\mathcal{D}}$.

Table 3: Power results summarized over 10 simulations. The best performing method is marked in bold font.

(a) Experiment A.

$v_1(x)$	D_Y	D_{XY}	KCE
$5x$	0.2	**1.0**	**1.0**
$0.5\cos(x)$	0.4	**1.0**	**1.0**
$0.1x^2$	0.0	**1.0**	0.9
$0.1\|x\|$	0.0	0.9	**1.0**
$0.1\max(0, 1-x)$	0.0	**1.0**	**1.0**
$0.1e^x$	0.0	0.8	**0.9**
$\frac{0.5}{x+3}$	0.2	**1.0**	**1.0**

(b) Experiment B.

$v_1(x)$	D_Y	D_{XY}	KCD
$2x$	**1.0**	0.9	**1.0**
$\cos(x)$	0.2	**1.0**	**1.0**
x^2	0.9	0.6	**1.0**
$x + 0.1\max(0, 1-x)$	0.0	**1.0**	**1.0**
$x + 0.1e^x$	0.1	**0.2**	**0.2**
$x + \frac{0.1}{x+3}$	**1.0**	**1.0**	**1.0**

(c) Experiment C.

$\mathcal{Y}$	H_A	D_Y	D_{XY}	KCD
$\mathbb{R}^p$	$\lambda = 0.8$	**1.0**	**1.0**	**1.0**
$\mathbb{R}^p$	$\lambda = 0.6$	0.8	**1.0**	0.9
$\mathbb{R}^p$	$\lambda = 0.4$	0.1	**0.4**	0.2
$\mathbb{R}^{p\times p}$	$Y_{ij} = X_i(X_j)^3$	**1.0**	0.7	**1.0**
$\mathbb{R}^{p\times p}$	$Y_{ij} = (X_i)^3(X_j)^3$	0.8	0.5	**0.9**
$\mathbb{R}^{p\times p}$	$Y_{ij} = \sin(X_i)\sin(X_j)$	0.1	0.1	**0.2**
$\mathcal{P}$	$\lambda = 0.8$	**1.0**	**1.0**	**1.0**
$\mathcal{P}$	$\lambda = 0.6$	0.6	**0.8**	**0.8**
$\mathcal{P}$	$\lambda = 0.4$	0.2	**0.3**	**0.3**