A Kernel Two-Sample Test with Selection Bias (Supplementary material)

Alexis Bellot[1,2] **Mihaela van der Schaar**[1,2,3]

[1]University of Cambridge, Cambridge, UK.
[2]Alan Turing Institute, London, UK
[3]University of California, Los Angeles, Los Angeles, USA

This appendix is outlined as follows.

- Section A proves all propositions and theorems.
  - Section A.1 proves Proposition 1.
  - Section A.2 proves Theorem 1.
  - Section A.3 proves Theorem 2.
  - Section A.4 proves Theorem 3.
- Section B gives a more detailed description and data generating mechanism of the example provided in the introduction.
- Section C gives a detailed description of other tests and our implementations.
- Section D discusses computational complexity and possible methods to speed up computations.

A PROOFS

In this section we prove the propositions and theorems described in the main body of this paper.

A.1 PROOF OF PROPOSITION 1

Assume that kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is characteristic (i.e. the mean embedding is injective) and that for all y, $w(x) > 0$ is bounded above by W. A kernel is called characteristic, if the maximum mean discrepancy between probability measures P_{Y^0} and P_{Y^1} induced by k is such that, $\text{MMD}(P_{Y^0}, P_{Y^1}) = 0$ if and only if $P_{Y^0} = P_{Y^1}$. (Gretton *et al.*, 2007) showed that Gaussian kernels are characteristic.

To prove the proposition we exploit the assumption $Y^0, Y^1 \perp\!\!\!\perp T|X$ and recover expectations with respect to the underlying random variables of interest (Y^0, Y^1). Assuming access to the propensity score, $e(x) = p(T = 1|X = x) = \mathbb{E}(I(T = 1)|X = x)$, and for any measurable function of our observed values Y, such as the kernel function k, we have that,

$$
\begin{aligned}
\mathbb{E}_{Y,Y^\star \sim P_{Y|T=1}}\left(\frac{k(Y,Y^\star)}{e(X)e(X^\star)}\right) &= \mathbb{E}_{Y,Y^\star}\left(\frac{TT^\star k(Y,Y^\star)}{e(X)e(X^\star)}\right) \\
&= \mathbb{E}_{Y,Y^\star}\left(\frac{I(T=1)I(T^\star=1)k(Y^1,Y^{1\star})}{e(X)e(X^\star)}\right) \\
&= \mathbb{E}_{X,X^\star}\left(\mathbb{E}_{Y,Y^\star}\left(\frac{I(T=1)I(T^\star=1)k(Y^1,Y^{1\star})}{e(X)e(X^\star)}|Y^1,Y^{1,\star},X,X^\star\right)\right) \\
&= \mathbb{E}_{Y^1,Y^{1,\star},X,X^\star}\left(\frac{k(Y^1,Y^{1\star})}{e(X)e(X^\star)}\mathbb{E}_{T,T^\star}\left(I(T=1)I(T^\star=1)|Y^1,Y^{1,\star},X,X^\star\right)\right) \\
&= \mathbb{E}_{Y^1,Y^{1,\star}}\left(k(Y^1,Y^{1,\star})\right)
\end{aligned}
$$

where recall that we use the notation y^1 for a realization of the random variable Y^1. I is the indicator function. This derivation shows that by taking weighted expectations with respect to the observed distribution $Y|T=1$ we can access expectations with respect to our distribution of interest Y^1. Similar derivations follow for data observed under $Y|T=0$ using the fact that $\mathbb{E}_{Y|T=0}(\frac{f(Y)}{1-e(X)}) = \mathbb{E}_{Y^0}(f(Y^0))$, for f any measurable function.

Now notice that the MMD(Y^0, Y^1) between Y^0 and Y^1 is defined in terms of expectations with respect to the random variables Y^0 and Y^1,

$$\text{MMD}(Y^0, Y^1) := \mathbb{E}_{Y^0,Y^{0,\star}} k(Y^0, Y^{0,\star}) + \mathbb{E}_{Y^1,Y^{1,\star}} k(Y^1, Y^{1,\star}) - 2\mathbb{E}_{Y^1,Y^0} k(Y^0, Y^1)$$

Thus with the above derivation we get that each term in the definition of WMMD$(Y|T=0, Y|T=1)$ is equal to each term in the definition of the MMD, which proves the proposition. □

A.2 PROOF OF THEOREM 1

Regularity conditions. The following notation is used in the statement on the regularity conditions of Theorem 1. Let $B_n = (b_{imn})$ and $W_n = (W_{ijn})$, for $i, j = 1, ..., n; n, m : 1, 2, ...$ Here W_n is a matrix of weights in $\mathbb{R}^{n\times n}$ and B_n is an orthogonal matrix in $\mathbb{R}^{m\times n}$ such that $B_n^T W_n B_n = \Lambda_n$, where Λ_n is a diagonal matrix with λ_{mn} as the m^{th} diagonal element. Assume $\lim_{n\to\infty} \lambda_{mn} = \lambda_m$ and let δ_{km} be the dirac delta function with $\delta_{km} = 1$ if $k = m$ and zero otherwise. Assume that the following regularity conditions hold,

1. $\max_{1\le i\le n} |b_{imn}| \to 0$ as $n \to \infty$ for each m.
2. $\sum_{i=1}^n b_{imn} b_{ikn} \to \delta_{mk}$ as $n \to \infty$ for all m, k.
3. $\sum_{i=1}^n \sum_{j=1}^n w_{ijn}^2 \to \sum_{m=1}^\infty \lambda_m^2 < \infty$.
4. $\sum_{i=1}^n \sum_{j=1}^n w_{ijn} b_{ikn} b_{jkn} \to \lambda_k$ as $n \to \infty$, for all m.

These conditions are sufficient by (De Wet *et al.*, 1973) for a square matrix of data-dependent weights $W = (w_i w_j)$ to be approximately diagonalizable, such that it admits an eigen-decomposition $B^T W B = \Lambda$.

Proof. Recall the definition of the empirical estimate of the WMMD^2,

$$\widehat{\text{WMMD}}^2 := \frac{1}{n(n-1)} \sum_{i\neq j: t_i = t_j = 1} w_i w_j k(y_i, y_j) + \frac{1}{m(m-1)} \sum_{i\neq j: t_i = t_j = 0} k(y_i, y_j) - \tag{1}$$

$$\frac{2}{nm} \sum_{i,j: t_i = 1, t_j = 0} w(x_i) k(y_i, y_j) \tag{2}$$

where the (y_i, t_i, x_i) are realization of the random variables (Y, T, X), and have assumed that n observations are made with $T = 1$ and m with $T = 0$. $w(x_i) = Pr(T_i = 1|X_i = x_i)/Pr(T_i = 0|X_i = x_i)$ is the density ratio giving the likelihood of an example i being observed under one population with respect to the other. We assume this ratio to be known (for now) and provide approximation bounds for our proposed approximation in Theorem 2 and 3. Our proof is presented in three parts, each one deriving the asymptotic behaviour of each one of the three terms in (1).

Note first that we may write the square integrable (centered) kernel k as a weighted sum of product of eigen-functions of the Hilbert-Schmidt operator defined by k (Gretton *et al.*, 2012),

$$k(y_i, y_j) = \sum_{k=1}^\infty \alpha_k \psi_k(Y_i) \psi_k(Y_j) \tag{3}$$

Consider now the first term in (1), it follows that,

$$\sum_{i=1}^n \sum_{j=1, j\neq i}^n w(x_i) w(x_j) k(y_i, y_j) = \sum_{i=1}^n \sum_{j=1, j\neq i}^n w_{ij} \sum_{k=1}^\infty \alpha_k \psi_k(Y_i) \psi_k(Y_j) \tag{4}$$

$$= \sum_{k=1}^\infty \alpha_k \sum_{i=1}^n \sum_{j=1, j\neq i}^n w_{ij} \psi_k(Y_i) \psi_k(Y_j) \tag{5}$$

where we have dropped the t_i's in the summation indices and have written $w_{ij} = w(x_i)w(x_j)$ for brevity. Using the degeneracy of k (in the sense that $\text{Var}[\mathbb{E}[k(y, y')]] = 0$), the eigen-functions $\psi_k(Y_i)$, $i = 1, ..., n$ are zero mean independent random variables by the independence of the Y_i. Using the above and the regularity conditions, Theorem 1 in (Verrill & Johnson, 1988) yields,

$$\frac{1}{n} \sum_{i=1}^n \sum_{j=1, j\neq i}^n w_{ij} \psi_k(Y_i) \psi_k(Y_j) \xrightarrow{d} \sum_{m=1}^\infty \lambda_m (Z_{km}^2 - 1) \tag{6}$$

where $Z_{km} \sim \mathcal{N}(0, 1)$ are $i.i.d.$.

The limiting distribution of the un-weighted term in (1) is that of a well-studied U-Statistic whose derivation can be found in Section 5.5.2 of (Serfling, 2009).

$$\frac{1}{m} \sum_{i=1: t_i = 0}^m \sum_{j=1, j\neq i: t_j = 0}^m k(Y_i, Y_j) \xrightarrow{d} \sum_{k=1}^\infty \alpha_k (V_k^2 - 1) \tag{7}$$

The limiting distribution of the cross term in (1) follows from a modification of the derivation of Theorem 1 in (De Wet *et al.*, 1973) and is given by,

$$\frac{1}{\sqrt{nm}} \sum_{i=1:t_i=1}^{n} \sum_{j=1:t_j=0}^{n} w'_{ij}\psi_k(Y_i)\psi_k(Y_j) \quad \xrightarrow{d} \quad \sum_{m=1}^{\infty} \lambda'_m Z_{km} V_{km} \tag{8}$$

where the eigenvalues (λ'_m) correspond to those of the eigen-decomposition of the weight matrix W' with $W'_{ij} = w(x_i)$ and where $V_{km} \sim \mathcal{N}(0,1)$ independently of $Z_{km} \sim \mathcal{N}(0,1)$. We prove (8) below.

We now combine these results. Define $t = m + n$, and assume $\lim_{m,n\to\infty} m/t = \rho_y$ and $\lim_{m,n\to\infty} n/t = \rho_x := (1 - \rho_y)$ for fixed $0 < \rho_x < 1$. Then,

$$t\widehat{\text{WMMD}}^2 \xrightarrow{d} \rho_x^{-1} \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda_m (Z_{km}^2 - 1) + \rho_y^{-1} \sum_{k=1}^{\infty} \alpha_k (V_k^2 - 1) - \frac{2}{\sqrt{\rho_x \rho_y}} \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda'_m Z_{km} V_{km} \tag{9}$$

In the case that both samples have equal size with total sample size n, we have that under $\mathcal{H}_0$,

$$n\widehat{\text{WMMD}}^2 \xrightarrow{d} \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda_m (Z_{km}^2 - 1) + \sum_{k=1}^{\infty} \alpha_k (V_k^2 - 1) - 2 \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda'_m Z_{km} V_{km} \tag{10}$$

The case of $P \neq Q$, under $\mathcal{H}_1$. The centered kernel k is non-degenerate since its expectation under assumption $\mathcal{H}_1$ is different from 0. The limiting distribution of WMMD can be derived by considering each term in the sum separately. For the first and third terms,

$$(\star) := \frac{1}{n(n-1)} \sum_{i \neq j : t_i = t_j = 1} w(x_i) w(x_j) k(y_i, y_j), \qquad (\star\star) := \frac{2}{mn} \sum_{i,j : t_i = 1, t_j = 0} w(x_i) k(y_i, y_j) \tag{11}$$

we get immediately by Theorem 2.1 from p. 4, (Shapiro *et al.*, 1979) that their limiting distributions are normal with mean $\mathbb{E}(\star)$ and variance $\text{Var}(\star)$, and mean $\mathbb{E}(\star\star)$ and variance $\text{Var}(\star\star)$, respectively. The middle term $\frac{1}{m(m-1)} \sum_{i \neq j : t_i = t_j = 0} k(y_i, y_j)$ is an un-weighted U-statistic whose limiting distribution is given by the results in section 5.5 (Serfling, 2009). As above, define $t = m + n$, and assume $\lim_{m,n\to\infty} m/t = \rho_y$ and $\lim_{m,n\to\infty} n/t = \rho_x := (1 - \rho_y)$ for fixed $0 < \rho_x < 1$. Collecting these results, we get under $\mathcal{H}_1$,

$$t^{1/2} \left(\widehat{\text{WMMD}}^2 - \text{WMMD}^2 \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right) \tag{12}$$

where we write $z = ((y_1, t = 1, x_1), (y_0, t = 0, x_0))$ for the joint sample under the two populations, and $h(z, z^\star) := w(x_1) w(x_1^\star) k(y_1, y_1^\star) + \mathbb{E} k(y_0, y_0^\star) - 2 w(x_1) k(y_1, y_0^\star)$. $\sigma^2_{\mathcal{H}_1} := \text{Var}_z \left(\mathbb{E}_{z^\star} h(z, z^\star) \right)$ (Serfling, 2009; Gretton *et al.*, 2012).

Proof of equation (8). The proof is a modification of the result of the convergence of degenerate U statistics on p. 761 in (Gretton *et al.*, 2012) and of the derivation of Theorem 1 in (De Wet *et al.*, 1973).

Consider,

$$T_k := \frac{1}{\sqrt{nm}} \sum_{i=1:t_i=1}^{n} \sum_{j=1:t_j=0}^{m} w'_{ij}\psi_k(Y_i)\psi_k(Y_j) \tag{13}$$

and define for each k,

$$w^*_{ij} := \sum_{s=1}^{S} \lambda_s b_{isk} b_{jsk}, \qquad T^*_k := \frac{1}{\sqrt{nm}} \sum_{i=1:t_i=1}^{n} \sum_{j=1:t_j=0}^{n} w'_{ij}\psi_k(Y_i)\psi_k(Y_j) \tag{14}$$

We will start by showing that $\sum_{i=1}^{n} \sum_{j=1}^{m} \left(w_{ij} - w^*_{ij} \right)^2 \to 0$ as $n, m \to \infty$. Note that this implies that $\text{Var}(T^*_k - T_k) \to 0$ and thus that the distributions of T^*_k and T_k coincide in the limit. We will proceed by showing first the convergence of the

sum of squares and then we derive the distribution of T_k^*. Using the definitions above, write,

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(w_{ij}-w_{ij}^{*}\right)^2=\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^2-2\sum_{s=1}^{S}\lambda_s\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}b_{isk}b_{jsk}+$$
$$\sum_{s=1}^{S}\sum_{t=1}^{S}\lambda_s\lambda_t\left(\sum_{i=1}^{n}b_{isk}b_{itk}\right)\left(\sum_{j=1}^{m}b_{jsk}b_{jtk}\right)$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^2-\sum_{s=1}^{S}\lambda_s^2-2\sum_{s=1}^{S}\lambda_s\left(\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}b_{iks}b_{jks}-\lambda_s\right)$$
$$+\sum_{s=1}^{S}\sum_{t=1}^{S}\lambda_t\lambda_s\left(\sum_{i=1}^{n}b_{isk}b_{itk}-\delta_{st}\right)\left(\sum_{j=1}^{m}b_{jsk}b_{jtk}-\delta_{st}\right)+$$
$$2\sum_{s=1}^{S}\lambda_s^2\left(\sum_{i=1}^{n}\sum_{j=1}^{m}b_{iks}^2-1\right) \tag{15}$$

where we have removed the group allocation indices t for clarity. Note here that the first and second term cancel each other by Assumption 1 of the regularity conditions, the third term is $\mathcal{O}(1)$ by Assumption 4 and the fourth and fifth terms are also $\mathcal{O}(1)$ by Assumption 2 and the properties of the dirac delta function.

Consider now T_k^* and rewrite it as,

$$T_k^*=\sum_{s=1}^{S}\lambda_s\left(\frac{1}{\sqrt{n}}\sum_{i=1:t_i=1}^{n}b_{isk}\psi_k(Y_i)\right)\left(\frac{1}{\sqrt{m}}\sum_{j=1:t_j=0}^{m}b_{jsk}\psi_k(Y_j)\right) \tag{16}$$

Define the length K vectors Ψ_n and Ψ'_m having k_{th} entries,

$$\Psi_{kn}=\left(\frac{1}{\sqrt{n}}\sum_{i=1:t_i=1}^{n}b_{isk}\psi_k(Y_i)\right),\qquad \Psi'_{km}=\left(\frac{1}{\sqrt{m}}\sum_{j=1:t_j=0}^{m}b_{jsk}\psi_k(Y_j)\right) \tag{17}$$

respectively. These have mean and covariance,

$$\mathbb{E}(\Psi_{kn})=0,\qquad \text{Cov}(\Psi_{kn},\Psi_{k'n})=\begin{cases}\frac{1}{m}\sum_{i=1}^{n}b_{isk}^2=1, & \text{if}\quad k=k'\\ 0, & \text{otherwise}\end{cases} \tag{18}$$

Moreover, the vectors Ψ_n and Ψ'_m are independent. The results (8) then holds by the Lindberg-Levy Central Limit Theorem (Serfling, 2009), Theorem 1.9.1A. □

A.3 PROOF OF THEOREM 2

We assume that for increasing sample size, as $n, m \to \infty$, we can approximate arbitrarily well the density ratio $w(x)$, for all x in our training data. This is justified by the following Lemma,

Lemma 1 (Lemma 1.4 (Gretton *et al.*, 2009a)) *Let $w(x_i) \in [0, B]$ be the optimal weight in the population sense, $Pr(T_i = 1|x_i) = w(x_i)Pr(T_i = 0|x_i)$. Assume we draw n samples from $X|T = 1$ and m samples from $X|T = 1$ independently and that $||\phi(x)|| \leq R$. Then, with probability at least $1 - \delta$,*

$$\left|\left|\frac{1}{n}\sum_{i=1:t_i=1}^{n}w(x_i)\phi(x_i)-\frac{1}{m}\sum_{j=1:t_j=0}^{m}\phi(x_j)\right|\right|\leq\left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}} \tag{19}$$

Note that because the optimization problem is convex the choice of $\hat{w}(x) := Pr(T_i = 1|x)/Pr(T_i = 0|x)$ uniquely minimizes the objective function with value 0, see Lemma 1.3, (Gretton *et al.*, 2009a). Thus by the argument above, we may

assume that for increasing sample size, as $n, m \to \infty$, $\hat{w}(x) \to w(x)$, for all x in the common support of the distributions $Pr(T_i = 1|x)$ and $Pr(T_i = 0|x)$.

Consider the first terms of $\widehat{\text{WMMD}}^2(;\hat{w})$ and $\widehat{\text{WMMD}}^2(;w)$, that denote the empirical WMMD^2 with estimated and true weights w respectively,

$$\hat{K}_{n,m} := \sum_{i=1:t_i=1}^{n} \sum_{j=1,j\neq i:t_j=1}^{m} \hat{w}_{ij} k(y_i, y_j), \quad \text{and} \quad K_{n,m} := \sum_{i=1:t_i=1}^{n} \sum_{j=1,j\neq i,t_j=1}^{m} w_{ij} k(y_i, y_j) \tag{20}$$

It holds that $\sum_{i=1}^{n} \sum_{j=1,j\neq i}^{m} \left(\hat{w}_{ij} - w_{ij}\right)^2 \to 0$ as $n, m \to \infty$ by the arguments at the end of section A.3. This implies that $\text{Var}(\hat{K}_{n,m} - K_{n,m}) \to 0$ and $E(|\hat{K}_{n,m} - K_{n,m}|^2) \to 0$ which means that $\hat{K}_{n,m} - K_{n,m}$ converges to 0 in L_2, and hence in distribution. The distributions of $\hat{K}_{n,m}$ and $K_{n,m}$ coincide in the limit.

The same derivations apply for the other two terms in the definition of $\widehat{\text{WMMD}}^2$. Therefore we conclude that $\widehat{\text{WMMD}}^2$ with estimated weights has the same asymptotic null and alternative distribution as $\widehat{\text{WMMD}}^2$ with known weights. In particular, asymptotically, its false positive rate is α and its power converges to 1.

A.4 PROOF OF THEOREM 3

We prove Theorem 3 by first stating and proving several Lemmas which bound the different terms of the inequality of interest.

Lemma 2 *In addition to the conditions of Lemma 1, assume there exists some $\hat{w}_i$, the empirical counterparts of the population weights estimated by matching kernel mean embeddings, such that,*

$$\left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \leq \epsilon \tag{21}$$

Then,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} w_i \phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} w(x_i) \phi(x_i) \right\| \leq \epsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}} \tag{22}$$

Proof. Note that by using Lemma 1 and the triangle inequality we immediately get,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{n} \sum_{i=1:t_i=1}^{n} w_i \phi(x_i) \right\| &\leq \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \\ &+ \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} w_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \\ &\leq \epsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}} \end{aligned} \tag{23}$$

□

Lemma 3. *Let $\widehat{WMMD}(w)$ be the weighted estimator of the MMD given i.i.d. distorted samples as defined in (1) with known (population) weights w, and similarly define $\widehat{WMMD}(\hat{w})$ with weights $\hat{w}$ estimated by matching the empirical kernel mean embeddings of the distorted samples. Then, given the conditions of Lemmas 1 and 2,*

$$\left| \widehat{\text{WMMD}}^2(\hat{w}) - \widehat{\text{WMMD}}^2(w) \right| \leq 2R(B+1) \left(\epsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}} \right) \tag{24}$$

Proof. Consider expanding the estimators,

$$\left|\widehat{\mathrm{WMMD}}^2(\hat{w}) - \widehat{\mathrm{WMMD}}^2(w)\right| = \frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j k(y_i,y_j) - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j k(y_i,y_j) \\ - \left(\frac{2}{nm}\sum_{i,j}\hat{w}_i k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} w(x_i)k(y_i,y_j)\right) \tag{25}$$

Note that the U-statistic in y cancel since these do not involve the weights.

First and second terms. We can bound the first and second terms as follows,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j k(y_i,y_j) - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j k(y_i,y_j) \tag{26}$$

$$= \frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j \left\langle\psi(y_i),\psi(y_j)\right\rangle - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j \left\langle\psi(y_i),\psi(y_j)\right\rangle \tag{27}$$

$$= \Big|\left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m \hat{w}_j\psi(y_j)\right\rangle \\ + \left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m w(x_j)\psi(y_j)\right\rangle\Big| \tag{28}$$

$$\leq \left|\left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m \hat{w}_j\psi(y_j)\right\rangle\right| \\ + \left|\left\langle \frac{1}{n}\sum_{i} \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i} w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m w(x_j)\psi(y_j)\right\rangle\right| \tag{29}$$

$$\leq 2BR\left(\epsilon + \left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}}\right) \tag{30}$$

where $\psi(y) := k(y,\cdot)$. Note that we have omitted the group allocation indices, these should be clear however from the i and j indices. The second equality follows by adding and subtracting $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j=1,j\neq i}^m w(x_i)\hat{w}_j\langle\psi(y_i),\psi(y_j)\rangle$ which factorizes into the given expression. The second to last inequality follows from the triangle inequality and the last inequality follows from the properties of norms and the results derived in Lemmas 1 and 2.

Third and fourth terms. The third and fourth terms (in brackets) are derived similarly and satisfy the following bounds,

$$\frac{2}{nm}\sum_{i,j}\hat{w}_i k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} w(x_i)k(y_i,y_j) \tag{31}$$

$$= \frac{2}{nm}\sum_{i,j}\hat{w}_i \left\langle\psi(y_i),\psi(y_j)\right\rangle - \frac{2}{nm}\sum_{i,j} w(x_i) \left\langle\psi(y_i),\psi(y_j)\right\rangle \tag{32}$$

$$= \left|\frac{1}{n}\sum_{i=1}^n \hat{w}_i\left\langle\psi(y_i), \frac{2}{m}\sum_{j=1}^m\psi(y_j)\right\rangle - \frac{1}{n}\sum_{i=1}^n w(x_i)\left\langle\psi(y_i), \frac{2}{m}\sum_{j=1}^m\psi(y_j)\right\rangle\right| \tag{33}$$

$$= \left|\left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{2}{m}\sum_{j=1}^m\psi(y_j)\right\rangle\right| \tag{34}$$

$$\leq 2R\left(\epsilon + \left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}}\right) \tag{35}$$

where the last inequality follows from the properties of norms and the results derived in Lemmas 1 and 2.

Finally, collecting the two bounds the lemma follows. □

Lemma 4. *Let $\widehat{WMMD}(w)$ be the weighted estimator of the MMD given $i.i.d.$ distorted samples as defined in (1) with known (population) weights w, and maximum kernel value R. Assume that $1 \leq w \leq B$ for all $x \in \mathcal{X}$. Then, with probability at least $1-\delta$,*

$$\left|\widehat{\text{WMMD}}^2(w) - \text{MMD}^2\right| \leq R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}} \tag{36}$$

where $m_2 := \lfloor m/2 \rfloor$.

Proof. Assuming the kernel $k(\cdot,\cdot)$ is bounded between 0 and R and the weights w bounded between 0 and B, we can infer function bounds such that $-2BR \leq wk(y_i, x_j) \leq R(B^2+1)$. By Theorem 10 in (Gretton *et al.*, 2012) which results from an application of the large deviation bound on U statistics due to Hoeffding we have that,

$$p\left(\left|\widehat{\text{WMMD}}^2(w) - \text{MMD}^2\right| > e\right) \leq \exp\left\{\frac{-2e^2 m_2}{R^2(B+1)^4}\right\} \tag{37}$$

Define $\delta = \exp\left\{\frac{-2e^2 m_2}{R^2(B+1)^4}\right\}$. Thus, with probability $1-\delta$,

$$\left|\widehat{\text{WMMD}}^2(w) - \text{MMD}^2\right| \leq R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}} \tag{38}$$

where $m_2 := \lfloor m/2 \rfloor$. □

We are ready to prove Theorem 3. This will be a straightforward combination of the lemmas given above.

Proof of Theorem 3. Let $\widehat{\text{WMMD}}(\hat{w})$ be the weighted estimator of the MMD given $i.i.d.$ distorted samples as defined in (1) with estimated weights $\hat{w}$. Assume conditions on Lemmas 1,2,3 and 4 above hold and that there exists an $\epsilon > 0$ such that,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{w}_i\phi(x_i) - \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\right\| \leq \epsilon \tag{39}$$

We may decompose the absolute difference between our weighted approximation using distorted samples and the population MMD as follows,

$$\begin{aligned}&\left|\widehat{\text{WMMD}}^2(\hat{w}) - \text{MMD}^2\right| \\ &\qquad \leq \left|\widehat{\text{WMMD}}^2(\hat{w}) - \widehat{\text{WMMD}}^2(w)\right| + \left|\widehat{\text{WMMD}}^2(w) - \text{MMD}^2\right|\end{aligned} \tag{40}$$

Then using Lemma 3 to bound the first term and Lemma 4 to bound the second term, we get that with probability at least $1-\delta$,

$$\begin{aligned}&|\widehat{\text{WMMD}}^2(\hat{w}) - \text{MMD}^2| \leq \\ &\qquad R(B+1)\left(2\epsilon + 2\left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}} + (B+1)\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}}\right)\end{aligned} \tag{41}$$

where $m_2 := \lfloor m/2 \rfloor$. □

B DETAILS ON THE INTRODUCTORY EXAMPLE

The example is used to illustrate the need for adjusting for confounding variables. For a total of 500 individuals we generated random education data X by sampling from a uniform distribution between 0 and 10, from which we derived the post-intervention income Y^0 and Y^1 by simply adding a standard random Gaussian noise variable to these values (in this case $\mathcal{H}_0$ holds: the distributions are equal). We generated male $T=1$ and female $T=0$ data, our two populations ($S=1$), by selectively removing with probability 0.5 females with education level higher than 5 ($Pr(T=0|X>5) \approx 0.33$), and removing with probability 0.5 males with education level lower than 5 ($Pr(T=0|X<5) \approx 0.66$). We end up with approximately 150 individuals in each group, males with higher education levels than females on average. Observe that the underlying generating process is the same in both populations, only the marginal distribution of the education level changes. As is natural, a two-sample test that overlooks the differences in education will reject the hypothesis of equal data generating process for the income.

C DESCRIPTION AND IMPLEMENTATION OF TESTS

C.1 KERNEL MEAN MATCHING

The idea in Kernel Mean Matching is to minimize the mean distance between a weighted data distribution $w(x)Pr(T = 0|X = x)$ and corresponding target data distribution $Pr(T = 1|X = x)$ in a reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ with feature map $\phi : D \to \mathcal{H}$, such that $k(x, y) :=< \phi(x), \phi(y) >$. Mean distance between these distributions is measured by computing the Maximum Mean Discrepancy between feature representations.

Under the assumption that $Pr(T = 1|X = x)$ is absolutely continuous with respect to $Pr(T = 0|X = x)$, i.e. $Pr(T = 1|X = x) = 0$ whenever $Pr(T = 0|X = x) = 0$, and that the RKHS kernel k is universal (see (Gretton *et al.*, 2012)) it has been shown that minimizing MMD with respect to the weights converges to $Pr(T = 1|X = x) = w(x)Pr(T = 0|X = x)$ (Gretton *et al.*, 2009a).

The optimization problem,

$$\underset{0<w<B}{\operatorname{argmin}} \|\mathbb{E}_{P_{X|T=0}} w(x)\phi(x) - E_{P_{X|T=1}} \phi(x)\|_{\mathcal{H}_K} \tag{42}$$

in finite samples reduces to the quadratic program,

$$\underset{0<w<B}{\operatorname{argmin}} \left|\left| \frac{1}{n} \sum_{i:t_i=0} w_i \phi(x_i) - \frac{1}{m} \sum_{i:t_i=1} \phi(x_i) \right|\right|^2 = \underset{0<w<B}{\operatorname{argmin}} \left(\frac{1}{n^2} w^T K w - \frac{2}{n} \kappa^T w + const. \right) \tag{43}$$

where K is a matrix with (i, j) entries $k(x_i, x_j)$ and $\kappa_i = \frac{n}{m} \sum_{j:t_j=1} k(x_i, x_j)$ as in (Gretton *et al.*, 2009a). n is the number of samples with $T = 0$ and m is the number of samples with $T = 1$.

This problem can then be solved with a convex optimization solver. We use `cvxpy` in python.

C.2 HYPERPARAMETER SELECTION FOR HIGH POWER

The population quantity WMMD $= 0$ if and only if the distributions under consideration are equal, for any choice of kernel hyperparameters. With finite sample size n, decisions must rely on inference based on the empirical WMMD, and some hyperparameters will give higher power than others. A popular strategy is to set the bandwidth σ of the Gaussian kernel to the median squared pairwise distance between input data, but can be sub-optimal when the scale of the difference between populations differs from the scale of the difference within populations themselves. Instead, we follow the approaches of (Sutherland *et al.*, 2016; Jitkrittum *et al.*, 2017) and choose σ so as to maximize the test power, i.e. the probability of rejecting $\mathcal{H}_1$ when it is false.

Proposition. (Approximate power of test statistic (Sutherland *et al.*, 2016)). *Under $\mathcal{H}_1$, for large n and fixed r, the test power $Pr(n\widehat{\text{WMMD}}^2 > r) \approx 1 - \Phi(\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} - \sqrt{n}\frac{\text{WMMD}^2}{\sigma_{\mathcal{H}_1}})$, where Φ denotes the cumulative distribution function of the standard normal distribution, and $\sigma_{\mathcal{H}_1}$ is defined as in Theorem 1.*

Assume that n is sufficiently large. Following the same argument as in (Jitkrittum *et al.*, 2017), in $\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} - \frac{\text{WMMD}^2}{\sigma_{\mathcal{H}_1}}$, we observe that the first term $\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} = \mathcal{O}(n^{-1/2})$ goes to 0 as $n \to \infty$ because $\sigma^2_{\mathcal{H}_1} = \mathcal{O}(n^{-1})$, while the second term, $\sqrt{n}\frac{\text{WMMD}^2}{\sigma_{\mathcal{H}_1}} = \mathcal{O}(n^{1/2})$, dominates the first one for large n. Thus, the parameters that maximize the test power are given by $\theta^* = \text{argmax}_\theta \; p(n\widehat{\text{WMMD}}^2 > r) \approx \frac{\text{WMMD}^2}{\sigma_{\mathcal{H}_1}}$. Since WMMD and $\sigma_{\mathcal{H}_1}$ are unknown, to maintain the validity of the hypothesis test we divide the sample into a training set, used to compute $\frac{\widehat{\text{WMMD}}^2}{\hat{\sigma}_{\mathcal{H}_1}}$ and choose the kernel, and a testing set used to perform the final hypothesis test with the learned kernel. The empirical estimate of the variance $\hat{\sigma}_{\mathcal{H}_1}$ that appears in our objective is approximated up to second order terms, similarly to (Sutherland *et al.*, 2016).

C.3 B-TEST: A MODIFICATION THAT USES PROPENSITY SCORES

An alternative to the weighted MMD test is a B-test (block-based test): the idea is to break the data into homogeneous blocks by stratifying subjects into mutually exclusive subsets based on their estimated propensity score. Recall that the propensity

score is defined as $e(x) := Pr(T = 1|X)$, the probability of group assignment given confounding variables. After this stage, we compute a two sample test statistic on each block, and average these quantities to obtain the test statistic.

More specifically, subjects are ranked according to their estimated propensity score and then stratified into subsets based on previously defined thresholds of the estimated propensity score. Because population assignment is essentially at random for individuals with the same propensity value, we expect mean comparisons within this group to be unbiased. (Rosenbaum & Rubin, 1983) showed that stratification based on the propensity score will balance x, in the sense that within strata homogeneous in $e(x) = Pr(T = 1|x)$, the distribution of x will be equal in the two populations.

For an individual block, laying on the main diagonal and starting at position $(i-1)B+1$, the statistic $\eta(i)$ is calculated as,

$$\eta(i) := \frac{1}{\binom{B}{2}} \sum_{a=(i-1)B+1}^{iB} \sum_{b=(i-1)B+1 \neq a}^{iB} h(y_{a,0}, y^{\star}_{b,0}, y_{a,1}, y^{\star}_{b,1}) \tag{44}$$

where $h(y_0, y^{\star}_0, y_1, y^{\star}_1) = k(y_0, y^{\star}_0) + k(y_1, y^{\star}_1) - k(y_0, y^{\star}_1) - k(y^{\star}_0, y_1)$, y_0 is a sample from $Y|T = 0$, y_1 a sample from $Y|T = 1$ and superscript $\star$ denotes an independent copy. The overall test statistic is then,

$$\eta = \frac{B}{n} \sum_{i=1}^{\frac{n}{B}} \eta(i) \tag{45}$$

The choice of B determines the accuracy of the balancing procedure and computation time - at one extreme is exact matching based on the propensity score and the linear-time MMD suggested by (Gretton *et al.*, 2012) where we have $n/2$ blocks of size $B = 2$, and at the other extreme is the unbalanced and usual full MMD with 1 block of size n. We chose as a default to divide both populations into $\sqrt{n}$ blocks as proposed in (Zaremba *et al.*, 2013).

B-test of (Zaremba *et al.*, 2013) assumes that $B \to \infty$ together with n, which implies that the statistic $\hat{\eta}$ defined in (45) under the null distribution satisfies,

$$\sqrt{nB}\hat{\eta} \to_d \mathcal{N}(0, 4\sigma^2) \tag{46}$$

where $\sigma^2 = E_{X,X'}(k(X, X')^2) + (E_{X,X'}k(X, X'))^2 - 2E_X[(E_{X'}k(X, X')^2]$ that can be estimated directly or by considering the empirical variance of the statistics computed within each of the blocks.

C.4 ANCOVA

Analysis of covariance (ANCOVA) are a general statistical procedure derived from a general linear model which blend ANOVA and regression. Conventionally, ANCOVA evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, that is confounders. In existing implementations (Tabachnick *et al.*, 2019) these suffer from a number of limitations such as the assumption of an underlying linear feature/outcome mapping and normality of residuals.

In our implementation we proceed as follows. We fit a Random Forest regression model on the confounding variables to approximate the outcome variable Y. Since in our experiments we consider Y to be multivariate, we fit a different regression model for each dimension of Y. We interpret the resulting residuals as being independent of confounders given group assignments and use those to proceed with testing. Because of the computational burden of this procedure, we fit the well-known Hotelling T^2 test (Hotelling, 1992) on the residuals to decide whether Y^0 and Y^1 share the same generating process up to confounding variables.

D COMPUTATIONAL COMPLEXITY

The computational complexity of the WMMD^2 is quadratic in the number of samples due to the need to compute the Kernel matrix, similarly to the plain implementation of the MMD^2. When permutations are chosen to approximate the null distribution, this procedure can be overly time consuming for large data sets. Below we briefly describe existing approximations that can be used with the WMMD^2 to speed up computations.

- Gamma approximation to the null (Gretton *et al.*, 2009b). This procedure consist of using a two-parameter Gamma distribution that we fit by matching the first and second moments of the empirical MMD^2. Such approximations can be accurate in practice and much faster, although they remain heuristics with no consistency guarantees.
- Linear time test (Gretton *et al.*, 2012). Another alternative would be to randomly subsample the data such as to make the computational complexity linear in the original number of samples. The drawback is that power is often overly reduced as a result.
- Kernel matrix approximation with low-dimensional random features (Rahimi & Recht, 2008). To accelerate the computation of the kernel matrix, one may map the input data to a randomized low-dimensional feature space and compute inner products based on these representations. (Rahimi & Recht, 2008) showed that by projecting unto a suitable basis the inner products of the transformed data are approximately equal to those in the feature space of a user specified shift-invariant kernel.

Computational complexity of weight estimation – The problem is convex and so in theory can be solved in polynomial time, in the worst case it scales cubically with the number of samples. However, complexity measures for quadratic solvers tend to be very conservative; in practice, we have found that far fewer iterations are needed than the theoretical bounds suggest. Our running times scaled approximately linearly with the number of samples.

REFERENCES

De Wet, T, Venter, JH, *et al.* 1973. Asymptotic distributions for quadratic forms with applications to tests of fit. *The Annals of Statistics*, **1**(2), 380–387.

Gretton, Arthur, Borgwardt, Karsten, Rasch, Malte, Schölkopf, Bernhard, & Smola, Alex J. 2007. A kernel method for the two-sample-problem. *Pages 513–520 of: Advances in neural information processing systems*.

Gretton, Arthur, Smola, Alex, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten, & Schölkopf, Bernhard. 2009a. Covariate shift by kernel mean matching. *Dataset shift in machine learning*.

Gretton, Arthur, Fukumizu, Kenji, Harchaoui, Zaid, & Sriperumbudur, Bharath K. 2009b. A fast, consistent kernel two-sample test. *Pages 673–681 of: Advances in neural information processing systems*.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, & Smola, Alexander. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, **13**(Mar), 723–773.

Hotelling, Harold. 1992. The generalization of Student's ratio. *Pages 54–65 of: Breakthroughs in statistics*. Springer.

Jitkrittum, Wittawat, Xu, Wenkai, Szabó, Zoltán, Fukumizu, Kenji, & Gretton, Arthur. 2017. A linear-time kernel goodness-of-fit test. *Pages 262–271 of: Advances in Neural Information Processing Systems*.

Rahimi, Ali, & Recht, Benjamin. 2008. Random features for large-scale kernel machines. *Pages 1177–1184 of: Advances in neural information processing systems*.

Rosenbaum, Paul R, & Rubin, Donald B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Serfling, Robert J. 2009. *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons.

Shapiro, Connie P, Hubert, Lawrence, *et al.* 1979. Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *The Annals of Statistics*, **7**(4), 788–794.

Sutherland, Dougal J, Tung, Hsiao-Yu, Strathmann, Heiko, De, Soumyajit, Ramdas, Aaditya, Smola, Alex, & Gretton, Arthur. 2016. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*.

Tabachnick, Barbara G, Fidell, Linda S, & Ullman, Jodie B. 2019. *Using multivariate statistics*. Vol. 7. Pearson Boston, MA.

Verrill, Steve, & Johnson, Richard A. 1988. Asymptotic distributions for quadratic forms with applications to censored data tests of fit. *Communications in Statistics-Theory and Methods*, **17**(12), 4011–4024.

Zaremba, Wojciech, Gretton, Arthur, & Blaschko, Matthew. 2013. B-test: A non-parametric, low variance kernel two-sample test. *Pages 755–763 of: Advances in neural information processing systems*.