Appendix: Two-sample Testing Using Deep Learning

A PROOF OF THEOREMS

A.1 Control of type-1 error rate

Proof of Theorem 3.1. (i) Under $p = q$, it holds that $\mathbb{E}[\phi(X_1)] = \mathbb{E}[\phi(Y_1)]$ and $\Sigma = \mathrm{Cov}(\phi(X_1)) = \mathrm{Cov}(\phi(Y_1))$. Then we have

$$\begin{aligned}
&\sqrt{\frac{nm}{n+m}} D_{n,m}(\phi) \\
&= \sqrt{\frac{nm}{n+m}} \left(\frac{1}{n} \sum_{i=1}^{n} \phi(X_i) - \mathbb{E}[\phi(X_i)] \right. \\
&\left. - \frac{1}{m} \sum_{i=1}^{m} \phi(Y_i) - \mathbb{E}[\phi(Y_i)] \right) \\
&= \sqrt{\frac{m}{n+m}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\phi(X_i) - \mathbb{E}[\phi(X_i)]) \\
&- \sqrt{\frac{n}{n+m}} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\phi(Y_i) - \mathbb{E}[\phi(Y_i)])
\end{aligned}$$

Then the first term in the last expression converges in distribution against $\mathcal{N}(0, r\Sigma)$ and the second term converges in distribution against $\mathcal{N}(0, (1-r)\Sigma)$ by a multivariate Central Limit Theorem (note that $\phi(X_1)$ lies within $[-1, 1]^H$ and hence all moments are finite). Since all X_i and Y_j are jointly independent, the limiting distributions are also independent, hence the whole term converges against $\mathcal{N}(0, r\Sigma) - \mathcal{N}(0, (1-r)\Sigma) = \mathcal{N}(0, \Sigma)$.

(ii) By *(i)* and the continuous mapping theorem, $S_{n,m}(\phi, \mathcal{X}_n, \mathcal{Y}_m) \xrightarrow{d} ||\zeta||^2$, where $\zeta \sim \mathcal{N}(0, \Sigma)$. Since Σ is positive semi-definite, there exist an orthogonal matrix Q and a diagonal matrix $L = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ such that $\Sigma = QLQ^\top$. Then we have

$$||Q\zeta||^2 = \zeta^\top Q^\top Q \zeta = \zeta^\top \zeta = ||\zeta||^2,$$

and $Q\zeta \sim \mathcal{N}(0, L)$, hence the claim.

(iii) By the weak law of large numbers, $\hat{\Sigma}_{n,m} \xrightarrow{p} \Sigma$, and hence by *(i)* and Slutsky's Theorem

$$\sqrt{\frac{nm}{n+m}} \hat{\Sigma}_{n,m}^{-\frac{1}{2}} D_{n,m}(\phi) \xrightarrow{d} \mathcal{N}(0, I).$$

The rest follows again by the continuous mapping theorem.

□

A.2 Proof of Consistency

Before we begin the proof we start with some auxiliary definitions and preliminary results.

As in Section 3.1 we can use the regression framework with $(Z_i, t_i)_i \subset \mathbb{R}^d \times \{-1, 1\}$. Then $Z_i | t_i = 1 \sim p$, $Z_i | t_i = -1 \sim q$ and similarly for (Z_i', t_i') and all jointly independent. As we assume $\Pr(t = 1) = \Pr(t = -1) = \frac{1}{2}$, the distribution of (Z, t) is fully determined by specifying p and q and hence we write for the expected value e.g. $\mathbb{E}_{p,q}[f(Z, t)]$ for some function f.

We define the loss function

$$L(t, \hat{t}) := 1 - t\hat{t} \in [0, 2]$$

with corresponding empirical and expected risks

$$\begin{aligned}
R_N'(\psi) &= \frac{1}{N} \sum_{i=1}^{N} 1 - t_i' \psi(Z_i'), \\
R'(\psi) &= 1 - \mathbb{E}_{p',q'}[t\psi(Z)].
\end{aligned}$$

The Bayes risk under the transfer task will be denoted as $R'^* = \inf_{f \in \mathcal{M}} R'(f) = 1 - \epsilon'$ where $\mathcal{M}$ is class of all Borel-measurable functions from $\mathbb{R}^d \to [-1, 1]$

Selecting ϕ_N is equivalent to (inexact) empirical risk minimization over $\mathcal{G}_N := \{w^\top \phi | \phi \in \mathcal{TF}_N, ||w|| \leq 1\}$, i.e.

$$R_N'(\psi_N) \leq \min_{\psi \in \mathcal{G}_N} R_N'(\psi) + \eta$$

where $\psi_N = w^\top \phi_N \in \mathcal{G}_N$ for some $||w_N|| \leq 1$.

The following Lemma is based on Theorem 1 in (Golowich et al., 2017) and we will need it to bound the complexity of the neural network function class $\mathcal{G}_N$.

Lemma A.1. *Let the data be a.s. be bounded by some $B > 0$ and*

$$\begin{aligned}
\mathcal{G} := \Big\{ & w^\top \tanh \circ W_{D'-1} \circ \sigma \circ \ldots \circ \sigma \circ W_1 : \mathbb{R}^d \to \mathbb{R} \Big| \\
& W_1 \in \mathbb{R}^{H \times d}, W_j \in \mathbb{R}^{H \times H} \text{ for } j = 2, \ldots, D'-1, \\
& w \in \mathbb{R}^H \text{ with } ||w|| \leq 1, \\
& \prod_{j=1}^{D'-1} ||W_j||_{Fro} \leq \beta, D' \leq D \Big\}
\end{aligned}$$

Then, the empirical Rademacher complexity of $\mathcal{G}$ can

be bounded as:

$$\hat{\mathcal{R}}_N(\mathcal{G}) \leq \frac{B(d+1)(\sqrt{2\log(2)(D-1)}+1)\beta}{\sqrt{N}}.$$

Proof of Lemma A.1. We define auxiliary function classes

$$\mathcal{G}^s_{D-1} := \Big\{ W_{D'-1} \circ \sigma \circ \ldots\ldots \sigma \circ W_1 : \mathbb{R}^d \to \mathbb{R}^s \Big| \\ W_1 \in \mathbb{R}^{H\times d}, W_j \in \mathbb{R}^{H\times H} \text{ for } j = 2,\ldots,D'-2, \\ W_{D'-1} \in \mathbb{R}^{s\times H}, \\ \prod_{j=1}^{D'-1} ||W_j||_{Fro} \leq \beta, D' \leq D \Big\}$$

for $s \in \{1, H\}$.

Then we can rewrite $\mathcal{G}$ as

$$\begin{aligned}\mathcal{G} &= \{\sum_{j=1}^{H} w_j \tanh\circ\phi_j | ||w|| \leq 1, \phi \in \mathcal{G}^H_{D-1}\} \\ &\subset \{\sum_{j=1}^{H} w_j \tanh\circ\phi_j | ||w|| \leq 1, \phi_j \in \mathcal{G}^1_{D-1}\} \\ &\subset \sum_{j=1}^{H} \{w \tanh\circ\phi | |w| \leq 1, \phi \in \mathcal{G}^1_{D-1}\}.\end{aligned}$$

Therefore we can bound the Rademacher complexity as

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{G}) &\leq H\hat{\mathcal{R}}_n(\{w\tanh\circ\phi | |w| \leq 1, \phi \in \mathcal{G}^1_{D-1}\}) \\ &\leq H\hat{\mathcal{R}}_n(\{\tanh\circ\phi | \phi \in \mathcal{G}^1_{D-1}\}) \\ &\leq H\hat{\mathcal{R}}_n(\mathcal{G}^1_{D-1})\end{aligned}$$

by standard learning theory arguments. For $\mathcal{G}^1_{D-1}$, we use the Rademacher bound found in (Golowich et al., 2017) Theorem 1 (we cannot use the Theorem directly on $\mathcal{G}$ since tanh is not positive homogeneous):

$$\hat{\mathcal{R}}_N(\mathcal{G}^1_{D-1}) \leq \frac{B(\sqrt{2\log(2)(D-1)}+1)\beta}{\sqrt{N}}.$$

The original Theorem 1 in (Golowich et al., 2017) holds for depth $D-1$ networks, but we allowed networks of lower depth. However, one can fill up the networks to depth $D-1$ with identity weight matrices and identity activation functions; inspection of the proof of the Theorem then shows that the claim still holds.

Since $H = d+1$, the claim follows. □

With these preliminary notions set up, we can proceed with the actual proof of consistency.

Proof of Theorem 3.2. We intend to show that $R_{n,m}(\psi_N)$ is asymptotically strictly smaller than 1; the divergence of the test statistic then follows easily. We will proceed in 5 steps. First, we split $R(\psi_N) - R'^*$ into transfer error, estimation error (of the transfer task) and approximation error (of the transfer task). Second, we show that the approximation error converges to zero (due to a Universal Approximation Theorem for deep networks); third we show that the estimation error is asymptotically bounded by η, using a learning theory bound on the Rademacher complexity of the neural network function class. This together implies that $R(\psi_N)$ and R'^* are $(\delta+\eta)$-close asymptotically. Fourth, we show that the $R_{n,m}(\psi_N) - R(\psi_N) \xrightarrow{p} 0$ and from this we finally deduce that the test statistics diverge to $+\infty$.

1. Splitting the terms We have

$$R(\psi_N) - R'^* = [R(\psi_N) - R'(\psi_N)] + [R'(\psi_N) - R'^*].$$

The first term is bounded as follows:

$$\begin{aligned}&|R(\psi_N) - R'(\psi_N)| \\ &\quad= |\mathbb{E}_{p,q}[t\psi_N(Z)] - \mathbb{E}_{p',q'}[t\psi_N(Z)]| \\ &\quad\leq \frac{||\psi_N||_\infty}{2}(||p-p'||_{L_1(\mu)} + ||q-q'||_{L_1(\mu)}) \\ &\quad\leq \delta,\end{aligned}$$

due to boundedness of ψ_N and requirement *(ii)*.

The second term can again be split:

$$\begin{aligned}&R'(\psi_N) - R'^* \\ &\quad= \left[R'(\psi_N) - \min_{\psi\in\mathcal{G}_N} R'(\psi)\right] + \left[\min_{\psi\in\mathcal{G}_N} R'(\psi) - R'^*\right].\end{aligned}$$

2. Convergence of $\min_{\mathcal{G}_N} R'(\psi) - R'^*$: Let $\hat{\mu}$ be the Borel measure of Z (not conditioned on t), i.e. $\Pr(Z \in A) = \hat{\mu}(A)$ for any $A \subset \mathbb{R}^d$ Borel. Following a similar argument as Lemma 30.2 in (Devroye et al., 2013) then yields the following. For any fixed $\epsilon > 0$, select a measurable function $h : \mathbb{R}^d \to [-1,1]$ such that $|R(h) - R^*| \leq \frac{\epsilon}{4}$, and a compact set $K \subset \mathbb{R}^d$ with $\hat{\mu}(K) \geq 1 - \frac{\epsilon}{8}$. Then, since compact-support continuous functions are dense in $L_1(\mu)$, there exists a continuous function $f : \mathbb{R}^d \to [-1,1]$ with

$$\mathbb{E}[|f(Z) - h(Z)|\mathbb{1}_{Z\in K}] \leq \frac{\epsilon}{4}.$$

From the universal approximation theorem for deep ReLU networks in (Hanin, 2017), there exists $N_0 \geq 1$ such that for all $N \geq N_0$ we can find a $\psi \in \mathcal{G}_N$ with

$$\sup_{z\in K} |f(z) - \psi(z)| \leq \frac{\epsilon}{4}.$$

Note that the Theorem in (Hanin, 2017) holds for ReLU-networks, but since tanh is invertible one can apply the universal approximation theorem on the first node in the last hidden layer, select the $w_N = [1, 0, \dots, 0]^\top$ and still get the universal approximation property.

Combining these yields, for N large enough,

$$\begin{aligned}
\min_{\psi\in\mathcal{G}_N} R'(\psi) - R'^* &\leq R'(\psi) - R'^* \\
&= \mathbb{E}_{p',q'}[-t\psi(Z) + th(Z)] + R'(m) - R'^* \\
&\leq \mathbb{E}[|\psi(Z) - h(Z)|\mathbb{1}_{Z\in K}] + 2\hat{\mu}(K^c) + \frac{\epsilon}{4} \\
&\leq \mathbb{E}[|\psi(Z) - f(Z)|\mathbb{1}_{Z\in K}] \\
&\quad + \mathbb{E}[|f(Z) - h(Z)|\mathbb{1}_{Z\in K}] + \frac{\epsilon}{2} \\
&\leq \epsilon.
\end{aligned}$$

Then, since $\epsilon > 0$ was arbitrary, and $\min_{\psi\in\mathcal{G}_N} R'(\psi) \geq R'^*$ we get $\min_{\mathcal{G}_N} R'(\psi) \to R'^*$ as $N \to \infty$.

3. Asymptotic closeness of $R'(\psi_N)$ and $\min_{\mathcal{G}_N} R'(\psi)$: We can first bound by standard arguments:

$$\begin{aligned}
&R'(\psi_N) - \min_{\psi\in\mathcal{G}_N} R'(\psi) \\
&= [R'(\psi_N) - R'_N(\psi_N)] + \left[R'_N(\psi_N) - \min_{\psi\in\mathcal{G}_N} R(\psi)\right] \\
&\leq \max_{\psi\in\mathcal{G}_N} [|R'(\psi) - R'_N(\psi)|] \\
&\quad + \left[\min_{\psi\in\mathcal{G}_N} R'_N(\psi) + \eta - \min_{\psi\in\mathcal{G}_N} R(\psi)\right] \\
&\leq 2 \max_{\psi\in\mathcal{G}_N} |R'(\psi) - R'_N(\psi)| + \eta \\
&= 2 \sup_{h\in\mathcal{H}_N} \left|\mathbb{E}[h(Z', t')] - \frac{1}{N}\sum_{i=1}^N g(Z'_i, t'_i)\right| + \eta
\end{aligned}$$

as $R'_N(\psi_N) \leq \min R'_N(\psi) + \eta$, where we define $\mathcal{H}_N := \{(z,t) \mapsto L(\psi(Z), t) | \psi \in \mathcal{G}_N\}$ as the conjunction of neural networks with the loss function. The first term can be bound with high probability by two-sided Rademacher inequalities:

$$\begin{aligned}
\Pr\Bigg(\sup_{h\in\mathcal{H}_N} &\left|\mathbb{E}[h(Z', t')] - \frac{1}{N}\sum_{i=1}^N g(Z'_i, t'_i)\right| \\
&\leq 2\hat{\mathcal{R}}_N(\mathcal{G}_N) + 6\sqrt{\frac{\log(4/\zeta)}{2N}}\Bigg) \\
&\geq 1 - \zeta
\end{aligned}$$

for any $\zeta > 0$. This complexity bound follows from Theorem 11.3 in (Mohri et al., 2018) if we insert the function class $\hat{\mathcal{H}} := \mathcal{H}_N \cup 2 - \mathcal{H}_N$ by noting that the loss function is 1-Lipschitz in both its arguments non-negative and bounded from above by 2.

Setting $\epsilon := 2\hat{\mathcal{R}}_N(\mathcal{G}_N) + 6\sqrt{\frac{\log(4/\zeta)}{2N}}$ then yields

$$\begin{aligned}
\Pr\left(\sup_{h\in\mathcal{H}_N} \left|\mathbb{E}[h(Z', t')] - \frac{1}{N}\sum_{i=1}^N g(Z'_i, t'_i)\right| > \epsilon\right) \\
\leq 4\exp\left(-\frac{N(\epsilon - 2\hat{\mathcal{R}}_N(\mathcal{G}_N))^2}{18}\right). \qquad (2)
\end{aligned}$$

But Lemma A.1 bounds the Rademacher complexity as

$$\begin{aligned}
\hat{\mathcal{R}}_N(\mathcal{G}_N) &\leq \frac{B(d+1)\left(\sqrt{2\log(2)(D_N - 1)} + 1\right)\beta_N}{\sqrt{N}} \\
&\leq C\frac{\sqrt{D_N}\beta_N}{\sqrt{N}} \qquad (3)
\end{aligned}$$

for some $C > 0$ and D_N large enough. Then

$$\begin{aligned}
N(\epsilon - 2\hat{\mathcal{R}}_N(\mathcal{G}_N))^2 &\geq N\epsilon^2 - 4N\hat{\mathcal{R}}_N(\mathcal{G}_N) \\
&\geq N\epsilon^2 - 4C\sqrt{N}\sqrt{D_N}\beta_N,
\end{aligned}$$

and the last term diverges to ∞ if $\frac{\beta_N^2 D_N}{N} \to 0$. Hence, the right-hand side in equation (2) converges to 0.

This shows that

$$\Pr(R'(\psi_N) - \min_{\psi\in\mathcal{G}_N} R'(\psi) \leq \epsilon + \eta) \to 1$$

for any $\epsilon > 0$, i.e. $R'(\psi_N)$ is asymptotically η-close to $\min_{\mathcal{G}_N} R'(\psi)$ (in probability).

4. $R_{n,m}(\psi_N) - R(\psi_N) \xrightarrow{p} 0$ Next we need to show that the empirical risk (over (Z,t), not (Z',t')) also is asymptotically smaller than 1.

We look at $\xi_{N,i} := t_i\psi_N(Z_i)$, which is a triangular array of random variables on $[-1, 1]$. We will use a weak law of large numbers for triangular arrays, see Theorem 2.2.11 in (Durrett, 2019). Both requirements in the Theorem are satisfied since $\xi_{N,i}$ is bounded, and hence we get

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N t_i\psi_N(Z_i) - \mathbb{E}[t\psi_N(Z)] \\
&\qquad = \frac{\sum_{i=1}^N \xi_{N,i} - N\mathbb{E}[\xi_{N,i}]}{N} \xrightarrow{p} 0,
\end{aligned}$$

or equivalently $R_{n,m}(\psi_N) - R(\psi_N) \xrightarrow{p} 0$.

But as shown above, $R(\psi_N)$ is δ-close to $R'(\psi_N)$ and $R'(\psi_N)$ is asymptotically η-close to R'^*; hence we get

$$\Pr(R(\psi_N) - R'^* \leq \epsilon + \delta + \eta) \to 1$$

for any $\epsilon > 0$, and therefore

$$\Pr(R_{n,m}(\psi_N) - R'^* \leq \epsilon + \delta + \eta) \to 1$$

5. Divergence of test statistics Define $M_N = 1 - R_{n,m}(\psi_N)$, then

$$\Pr(M_N \geq \epsilon^* - \delta - \eta - \epsilon) \to 1$$

for any $\epsilon > 0$. Since $\delta + \eta < \epsilon^*$, we then have for any $r > 0$:

$$\Pr\left(\sqrt{\frac{nN}{m}} M_N > r\right) = \Pr\left(M_N > \sqrt{\frac{m}{nN}} r\right) \to 1,$$

i.e. $M_N \xrightarrow{p} +\infty$, since $\sqrt{\frac{m}{nN}} \to 0$.

Next, define

$$\hat{S}_{n,m} = \frac{nm}{n+m}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_N(X_i) - \frac{1}{m}\sum_{i=1}^{m}\psi_N(Y_i)\right),$$

i.e. the version of $S_{n,m}$ where the last layer is still selected on the training data instead of the test data. Then it holds that

$$\begin{aligned}
&\left|\sqrt{\frac{m}{n(m+n)}}\hat{S}_{n,m} - M_N\right| \\
=&\left|\frac{1}{m+n}\left(\frac{m}{n}\sum_{i=1}^{n}\psi_N(X_i) - \sum_{i=1}^{m}\psi_N(Y_i)\right)\right. \\
&\qquad \left. - \frac{1}{m+n}\sum_{i=1}^{m+n} t_i \psi_N(Z_i)\right| \\
=&\left|\frac{1}{m+n}\sum_{i=1}^{n}\psi_N(X_i)\right|\left|\frac{m}{n} - 1\right| \\
\leq&\left|\frac{m}{n} - 1\right| \to 0,
\end{aligned}$$

since all $|\psi_N(X_i)| \leq 1$ and $\frac{m}{n} \to 1$.

Hence, we also get

$$\Pr(\hat{S}_{n,m} > r) \to 1$$

for any $r > 0$. But

$$\begin{aligned}
S_{n,m} &= \frac{nm}{n+m}\left|\left|\overline{\phi_N(\mathcal{X}_n)} - \overline{\phi_N(\mathcal{Y}_m)}\right|\right|^2 \\
&= \frac{nm}{n+m}\sup_{||w||\leq 1} w^\top\left(\overline{\phi_N(\mathcal{X}_n)} - \overline{\phi_N(\mathcal{Y}_m)}\right) \\
&\geq \hat{S}_{n,m}
\end{aligned}$$

For the DFDA test statistic, we have

$$\begin{aligned}
T_{n,m} &= \frac{nm}{n+m} D_{n,m}^\top \hat{\Sigma}_{n,m}^{-1} D_{n,m} \\
&\geq \frac{nm}{n+m}||D_{n,m}||^2 \lambda_{\min}(\hat{\Sigma}_{n,m}^{-1}) \\
&= S_{n,m}\lambda_{\max}(\hat{\Sigma}_{n,m})^{-1}.
\end{aligned}$$

$\lambda_{\max}(\hat{\Sigma}_{n,m})$ is always positive (due to the $\rho_{n,m} > 0$ summand), and also bounded from above by some $C > 0$ (due to the boundedness of all individual entries), therefore $T_{n,m} \geq C^{-1} S_{n,m}$.

Hence we also have $S_{n,m}, T_{n,m} \xrightarrow{p} +\infty$. □

B DISTRIBUTIONS WITH UNBOUNDED SUPPORT

Considering the case where p' and q' have unbounded support, but requirements *(ii)*, *(iii)* and a variant of *(i)* in Theorem 3.2 are still satisfied, we can still prove a similar consistency result.

In particular, we can make p' and q' vary with N by replacing them with truncated, bounded-support versions that converge towards the true densities slowly enough. First, select p'_N and q'_N with support on $[-B_N, B_N]^d$ for some sequence $B_n \uparrow +\infty$, and $||p'_N - p'||_{L_1(\mu)} \to 0$ and $||q'_N - q'||_{L_1(\mu)} \to 0$. Then there exists a $N_0 > 0$ such that for all $N \geq N_0$, requirements *(i)*, *(ii)* and *(iii)* are satisfied for p'_N and q'_N. In practice these truncated variables can be achieved for example by rejection sampling from p' and q'.

The only part in the proof of Theorem 3.2 where we need the boundedness assumption on p' and q' is when bounding the Rademacher complexity of the class $\mathcal{G}_N$ in equation 3. The modified Rademacher bound now is

$$\begin{aligned}
\hat{\mathcal{R}}_N(\mathcal{G}_N) &\leq \frac{B_N(d+1)\left(\sqrt{2\log(2)(D_N - 1)} + 1\right)\beta_N}{\sqrt{N}} \\
&\leq C\frac{\sqrt{D_N} B_N \beta_N}{\sqrt{N}}.
\end{aligned}$$

The requirement for the exponent in equation (2) to diverge then is

$$\begin{aligned}
\frac{B_N^2 \beta_N^2 D_N}{N} &\to 0 \text{ instead of} \\
\frac{\beta_N^2 D_N}{N} &\to 0.
\end{aligned}$$

The rest of the proof is as before. We can summarize this as follows:

Theorem B.1. *Let $p \neq q$, $n = n', m = m'$ with $\frac{n}{m} \to 1$, $N = n + m$, $R'^* = 1 - \epsilon'$ the Bayes error for the transfer task with $\epsilon' > 0$. Furthermore, let $||p'_N - p'||_{L_1(\mu)} \to 0$ and $||q'_N - q'||_{L_1(\mu)} \to 0$ for sequences of μ-densities $(p'_N)_N$ and $(q'_N)_N$,*

Assume that the following holds:

(i) $\frac{B_N^2 \beta_N^2 D_N}{N} \to 0$, $B_N \to \infty$, $\beta_N \to \infty$ *and* $D_N \to \infty$ *for* $N \to \infty$,

(ii) $||p - p'||_{L_1(\mu)} + ||q - q'||_{L_1(\mu)} < 2\delta$,

(iii) $0 \leq \delta + \eta < \epsilon'$, *where* $\eta \geq 0$ *is the leniency parameter in training the network, and*

(iv) for each N, p'_N and q'_N have support on $[-B_N, B_N]^d$.

Then, as $N \to \infty$ both test test statistics $S(\phi_N, \mathcal{X}_n, \mathcal{Y}_m)$ and $T(\phi_N, \mathcal{X}_n, \mathcal{Y}_m)$ diverge in probability towards infinity, i.e. for any $r > 0$

$$\Pr\left(S(\phi_N, \mathcal{X}_n, \mathcal{Y}_m) > r\right) \to 1 \text{ and}$$
$$\Pr\left(T(\phi_N, \mathcal{X}_n, \mathcal{Y}_m) > r\right) \to 1.$$

C DIMENSIONALITY REDUCTION

In practice, we oftentimes first apply a PCA transformation on the data before computing the DFDA test statistic. Since we fit the PCA on the test data itself, however, the observations are not independent anymore and Theorem 3.1 is not directly applicable anymore. As an unsupervised linear transformation, however, we can show via a Slutsky-type argument that the normal approximation is still valid.

Theorem C.1. *Let $(\xi_i)_i$ and $(\xi'_i)_i$ be all jointly independent and identically distributed on $\mathbb{R}^d$ with bounded support and assume that $\frac{n}{n+m} \to r \in (0,1)$ as $n, m \to \infty$.*

Let $A_N \in \mathbb{R}^{s \times d}$ be a PCA transform, fitted on $\xi_1, \dots, \xi_n, \xi'_1, \dots, \xi'_m$ ($N = m + n$), for some $s \in \{1, \dots, d\}$. Let $\Sigma = Cov(\xi_1)$ with eigenvalues $\lambda_1, \dots, \lambda_d$ sorted in descending order, and assume that $\lambda_s \neq \lambda_{s+1}$ (if $s < d$).

Then

$$\sqrt{\frac{nm}{n+m}} \left(\frac{1}{n} \sum_{i=1}^{n} A_N \xi_i - \frac{1}{m} \sum_{i=1}^{m} A_N \xi'_i \right) \xrightarrow{d} \mathcal{N}(0, \Sigma')$$

as $n, m \to \infty$, where $\Sigma' = diag(\lambda_1, \dots, \lambda_s)$.

Note that the $\lambda_s \neq \lambda_{s+1}$ assumption is only necessary for uniqueness of the limiting distribution – in practice one can ignore this requirement.

Proof of Theorem C.1. Since A_N is a PCA transformation, A_N is the matrix with the normalized eigenvectors corresponding to the s largest eigenvalues of the empirical covariance matrix $\tilde{\Sigma}_{n,m}$. But, due to a weak law of large numbers, $\tilde{\Sigma}_{n,m} \xrightarrow{p} \Sigma$ and accordingly $A_N \xrightarrow{p} A$ with the population PCA A being the normalized eigenvectors corresponding to the s largest eigenvalues of Σ (without loss of generality we can assume the row-wise signs to be determined by some deterministic procedure, and hence for large enough N, A_N and A unique e.g. by requiring that the first non-zero entry in the vector be positive).

Due to the same argument as in the proof of Theorem 3.1 *(i)*,

$$\sqrt{\frac{nm}{n+m}} \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i - \frac{1}{m} \sum_{i=1}^{m} \xi'_i \right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Due to a multivariate Slutsky theorem, then

$$\begin{aligned}
&\sqrt{\frac{nm}{n+m}} \left(\frac{1}{n} \sum_{i=1}^{n} A_N \xi_i - \frac{1}{m} \sum_{i=1}^{m} A_N \xi'_i \right) \\
&\quad = A_N \sqrt{\frac{nm}{n+m}} \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i - \frac{1}{m} \sum_{i=1}^{m} \xi'_i \right) \\
&\quad \xrightarrow{d} \mathcal{N}(0, A \Sigma A^\top).
\end{aligned}$$

But as A consists of the orthogonal eigenvectors of Σ in descending order of eigenvalues, $A \Sigma A^\top = \Sigma'$.

□

D ADDITIONAL EMPIRICAL ANALYSIS

D.1 Parameters for SCF and ME tests

For the SCF and ME test, hyperparameters have to be chosen, namely the number of locations/frequencies at which to test, the kernel-selection strategy and whether to optimize over the frequencies/locations or to use a simple heuristic. We found that if the number of locations/frequencies J is chosen too large, the tests oftentimes strongly violate the significance level. Hence, we grow J with the number of samples according to what still gives reasonable type-1 error rates.

AM Audio Data Here we use the 'full' version of the parameter selection from (Jitkrittum et al., 2016) for both tests. Number of frequencies/locations were set to $J = 1$ when $m \in [10, 50]$, $J = 3$ for $m \in [75, 150]$ and $J = 10$ for $m \in [200, 1000]$.

Aircraft, Dogs and Birds Data For SCF we found the random location initialization without kernel optimization (and hence without data split) to work best. For ME, due to the high dimensionality, we selected the 'grid' version of the parameter optimization; the 'full' version did not seem to give considerable improvements above this. For the Aircraft and Dogs data, we selected $J = 1$ frequencies/locations for $m \in [10, 50]$ and $J = 3$ for $m \in [50, 200]$. For the Birds data we always use $J = 1$ ($m \in [10, 60]$).

Facial Expression Data Again we use random locations for SCF and grid-search kernel width for ME. For SCF, we fix $J = 1$ for all $m \in [10, 200]$. For ME, we choose $J = 1$ for $m \in [10, 50]$, $J = 3$ for $m \in [75, 100]$ and $J = 10$ for $m \in [150, 200]$.

D.2 Image Preprocessing

For the deep learning-based methods (DFDA, DMMD & C2ST), before evaluation, all image data is rescaled

to (224, 224) and normalized according to the requirements of the neural network.

For kernel-based tests we found different strategies to work differently well on each data set. Hence, for the Aircraft, Stanford Dogs and Birds data set, data is rescaled to (48, 48) dimensions and converted to grayscale. For the facial expression data, images were first cropped to the center (resulting in (462, 462) dimensions) and then rescaled to (96, 96) dimensions; no conversion to grayscale was performed. We found no increase in power for higher resolution (e.g. (224, 224)).

D.3 Sensitivity to Significance Level

Special care has to be taken if several hypotheses are tested at the same time, leading to a *multiple testing problem*. One simple approach to control the so-called familywise error rate (FWER, (Lehmann and Romano, 2006)), i.e., the probability of at least one wrong rejection of a null hypothesis, is the Bonferroni correction (Lehmann and Romano, 2006). The Bonferroni correction divides the original significance level α by the number of tests to be performed. Therefore, in many practical settings the significance level for each test will be considerably lower than the "standard" values of 0.05 or 0.01. This represents a problem in practice, since approximating the distribution in the tails usually is more challenging. Here we only give results for the asymptotic DFDA distribution, since permutation-based methods do not scale well to very low significance levels. Figure 3a shows that our method controls type-1 error rate at significance levels $5 \cdot 10^{-u}$ for $u = 2, 3, 4, 5$; Figure 3b shows that even at small significance levels, the DFDA can still maintain relatively high power.

D.4 Control of Type-1 Error Rate

Figure 4 shows that both DMMD and DFDA properly control the type-1 error rate even at low sample sizes.

D.5 Birds Experiments

Here we report results on another fine-grained classification data set, the Caltech-UCSD Birds-200-2011, Caltech-UCSD Birds-200-2011 (Wah et al., 2011). We selected two visually very similar species of birds, namely the "Blue-winged Warbler" and the "Hooded Warbler" for differentiation. Results are shown in Figure 5.

D.6 AM Audio Experiments

Data preprocessing consists of sampling the original audio signal at 8kHz, the resulting AM signal is sampled at 120kHz, and snippets of length 1000 are used for identification. Gaussian noise with standard deviation 1 is added to the samples after processing.

The model has four one-dimensional convolutional layers, each followed by Batch normalization, a ReLU activation and max-pooling. The last layer is fully connected, but only used for training the network, i.e., the feature extraction is fully convolutional. In contrast to the M5 network, we use an input layer with kernel size of 20 instead of 80 and the final global average pooling layer can be removed, to accommodate the significantly smaller input dimension of the audio snippets. We train the network to classify noisy AM snippets from the remaining songs on the album, with a multi-class cross-entropy loss and a L_2-regularization of 10^{-4} on all weights; we use the Adam optimizer for this task Kingma and Ba (2014).

D.7 Stanford Dogs Experiments

Table 2 shows the convolutional autoencoder architecture used in the experiments on the Stanford Dogs data set. The autoencoder was trained to optimize multi-scale structural similarity between input and output images.

The supervised training was performed with a network with the same encoder as in Table 2 and a fully connected layer on top, to classify the remaining 118 dog breeds. Again, we use the multi-class cross-entropy loss.

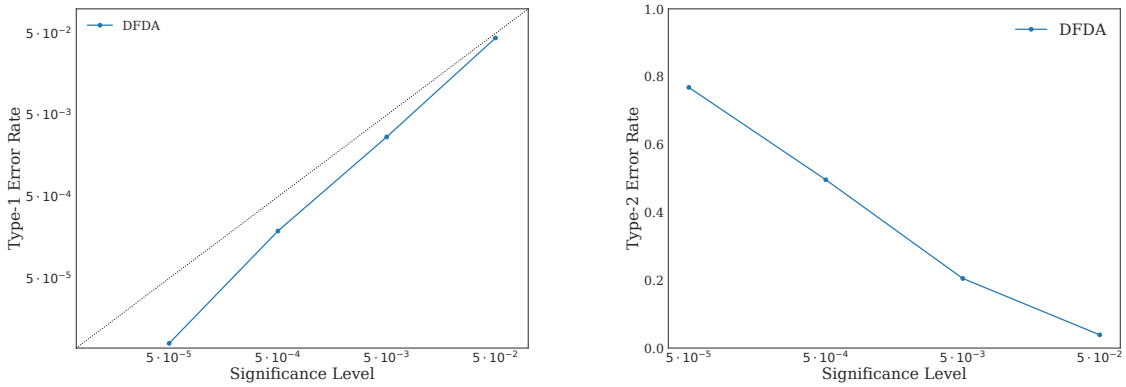
For both the supervised and the unsupervised task we use the Adam optimizer and L_2 regularization of size 10^{-4}.

D.8 KDEF Experiments

Note that Jitkrittum et al. (2016) and Lopez-Paz and Oquab (2016) only compared tests that use train/test splits. Hence, results therein are reported for n_{te}, which is the size of the test set of each sample, i.e. $n_{te} = \frac{1}{2}m$ in our case ($n_{te} = 201$ corresponds to $m = 402$).

D.9 Imagenet Training

For the aircraft, facial expression, and birds data set we use a ResNet-152, trained on the whole ILSVRC 2012 data set. Instead of training this network ourselves, we use the parameters and implementation provided in the PyTorch deep learning library Paszke et al. (2017).



(a) Type-1 error rate at low significance levels, $m = 50$. (b) Type-2 error rate at low significance levels, $m = 50$.

Figure 3: Results on the AM audio data for $m = 50$ with small significance levels α. We show average values over 10^6 tests, where we fixed the sample size m per population to be equal to 50. (a) Empirical type-1 error rates for small α values consistently lie below the expected type-1 error rate (dotted line). (b) Empirical type-2 error rates.

D.10 MRI Scan Preprocessing and Experiments

The T1 MRI scans acquired through the MP-RAGE protocol were selected from GSP and ADNI. The scans were standardized to $(256, 256, 256)$ and cropped to $(96, 96, 96)$ dimensions with isotropic voxels of 1mm. Model architecture is shown in table 3. The model was trained for 400 epochs on 1413 MRI scans from GSP. The loss function was set to the mean squared error and the batch size was set to one. No MRI scans from ADNI was used for training.

In our experiments, the *APOE* gene was used since it is known to be a risk factor for Alzheimer's disease; in practice, when one does not know which locus to test, a multistep-approach such as the one developed by Mieth et al. (2016) can be used to create a selection of candidate loci.

E CODE AND DATA

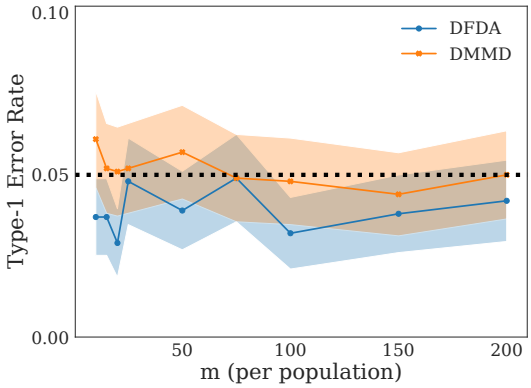
We provide an implementation of our methods at `https://github.com/mkirchler/deep-2-sample-test`.

All 2D imaging and audio data can be downloaded from the following sources:

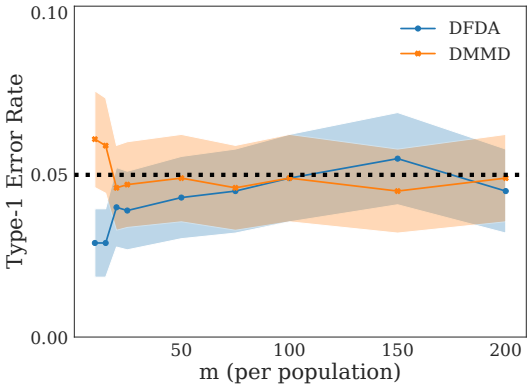
- Audio data: `http://dl.lowtempmusic.com/Gramatik-TAOR.zip`
- Aircraft data: `http://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/archives/fgvc-aircraft-2013b.tar.gz`
- Facial Expression data: `http://kdef.se/index.html`
- Stanford Dogs data: `http://vision.stanford.edu/aditya86/ImageNetDogs/images.tar`
- Birds data: `http://www.vision.caltech.edu/visipedia-data/CUB-200-2011/CUB_200_2011.tgz`

For MRI imaging data access to data has to be granted by the releasing institutions, see

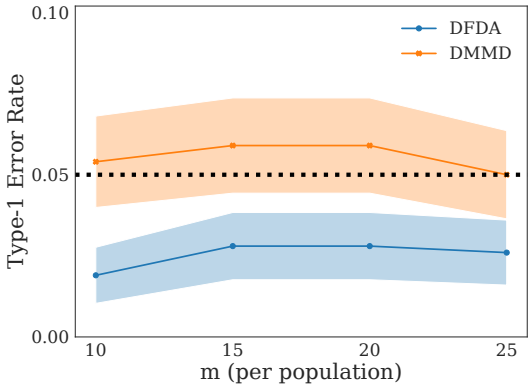
- GSP: `https://www.neuroinfo.org/gsp`
- ADNI: `http://adni.loni.usc.edu/data-samples/access-data/`



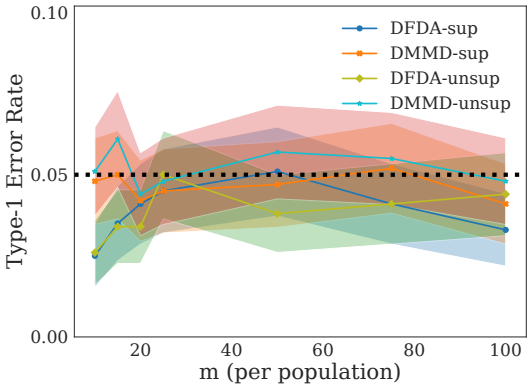
(a) Type-1 error rate on Aircraft data set.



(b) Type-1 error rate on facial expression data set.



(c) Type-1 error rate on Birds data set.



(d) Type-1 error rate on Stanford Dogs data set.

Figure 4: Empirical control of type-1 error rate on vision data sets.

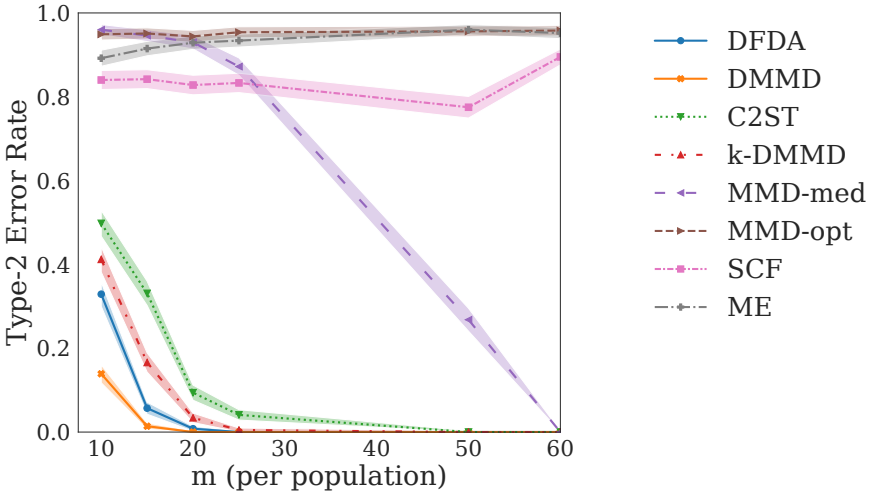


Figure 5: Type-2 error rate on Birds data set.

Table 2: Architecture of the convolutional autoencoder used for the Stanford Dogs experiments. For Conv and ConvTranspose layers, $[3 \times 3, f]$ denotes f 3×3 filters. Activation functions are always ReLUs except for the last convolutional layer (tanh) and the last ConvTranspose layer (sigmoid). After each Conv and ConvTranspose operation, a BatchNorm (Ioffe and Szegedy, 2015) operation was used. The output of the encoder part was used as feature map in our tests.

Input: $(3, 224, 224)$ image	
Encoder	
Conv	$[3 \times 3, 40]$
MaxPool	$[2 \times 2]$
Conv	$[3 \times 3, 80]$
MaxPool	$[2 \times 2]$
Conv	$[3 \times 3, 160]$
MaxPool	$[2 \times 2]$
Conv	$[3 \times 3, 240]$
MaxPool	$[2 \times 2]$
Conv	$[3 \times 3, 360]$
MaxPool	$[2 \times 2]$
Conv	$[3 \times 3, 2048]$
MaxPool	$[2 \times 2]$
Decoder	
ConvTranspose	$[3 \times 3, 360]$
Upsample	$[2 \times 2]$
ConvTranspose	$[3 \times 3, 240]$
Upsample	$[2 \times 2]$
ConvTranspose	$[3 \times 3, 160]$
Upsample	$[2 \times 2]$
ConvTranspose	$[3 \times 3, 80]$
Upsample	$[2 \times 2]$
ConvTranspose	$[3 \times 3, 40]$
Upsample	$[2 \times 2]$
ConvTranspose	$[3 \times 3, 3]$

Table 3: Architecture of the 3D convolutional autoencoder for the MRI data. For Conv and ConvTranspose layers, $[3 \times 3 \times 3, s, f]$ denotes f $3 \times 3 \times 3$ filters with strides of s. Activation functions are always ReLUs except for the last convolutional layer (linear). All convolutional operations are done without padding. The output of the encoder (1024 dimensions) is used as feature map in our tests.

Input: $(96, 96, 96)$ MRI scan	
Encoder	
Conv	$[3 \times 3 \times 3, 1, 8]$
Conv	$[2 \times 2 \times 2, 2, 16]$
Conv	$[3 \times 3 \times 3, 1, 32]$
Conv	$[2 \times 2 \times 2, 2, 64]$
Conv	$[2 \times 2 \times 2, 2, 128]$
Conv	$[2 \times 2 \times 2, 2, 256]$
Conv	$[2 \times 2 \times 2, 2, 256]$
Dense	
Decoder	
Dense	
Conv	$[3 \times 3 \times 3, 1, 256]$
ConvTranspose	$[2 \times 2 \times 2, 2, 256]$
ConvTranspose	$[2 \times 2 \times 2, 2, 128]$
ConvTranspose	$[2 \times 2 \times 2, 2, 64]$
ConvTranspose	$[2 \times 2 \times 2, 2, 32]$
Conv	$[3 \times 3 \times 3, 1, 16]$
ConvTranspose	$[2 \times 2 \times 2, 2, 8]$
Conv	$[3 \times 3 \times 3, 1, 1]$