Two-sample Testing Using Deep Learning

Matthias Kirchler[1,2] **Shahryar Khorasani**[1] **Marius Kloft**[2,3] **Christoph Lippert**[1,4]
[1]Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
[2]Technical University of Kaiserslautern, Germany
[3]University of Southern California, Los Angeles, United States
[4]Hasso Plattner Institute for Digital Health at Mount Sinai, New York, United States

Abstract

We propose a two-sample testing procedure based on learned deep neural network representations. To this end, we define two test statistics that perform an asymptotic location test on data samples mapped onto a hidden layer. The tests are consistent and asymptotically control the type-1 error rate. Their test statistics can be evaluated in linear time (in the sample size). Suitable data representations are obtained in a data-driven way, by solving a supervised or unsupervised transfer-learning task on an auxiliary (potentially distinct) data set. If no auxiliary data is available, we split the data into two chunks: one for learning representations and one for computing the test statistic. In experiments on audio samples, natural images and three-dimensional neuroimaging data our tests yield significant decreases in type-2 error rate (up to 35 percentage points) compared to state-of-the-art two-sample tests such as kernel-methods and classifier two-sample tests.*

1 INTRODUCTION

For almost a century, statistical hypothesis testing has been one of the main methodologies in statistical inference (Neyman and Pearson, 1933). A classic problem is to validate whether two sets of observations are drawn from the same distribution (null hypothesis) or not (alternative hypothesis). This procedure is called *two-sample test*.

Two-sample tests are a pillar of applied statistics and a standard method for analyzing empirical data in the sciences, e.g., medicine, biology, psychology, and social sciences. In machine learning, two-sample tests have been used to evaluate generative adversarial networks (Bińkowski et al., 2018), to test for covariate shift in data (Zhou et al., 2016), and to infer causal relationships (Lopez-Paz and Oquab, 2016).

There are two main types of two-sample tests: parametric and non-parametric ones. Parametric two-sample tests, such as the Student's t-test, make strong assumptions on the distribution of the data (e.g. Gaussian). This allows us to compute p-values in closed form. However, parametric tests may fail when their assumptions on the data distribution are invalid. Non-parametric tests, on the other hand, make no distributional assumptions and thus could potentially be applied in a wider range of application scenarios. Computing non-parametric test statistics, however, can be costly as it may require applying re-sampling schemes or computing higher-order statistics.

A non-parametric test that gained a lot of attention in the machine-learning community is the kernel two-sample test and its test statistic: the maximum mean discrepancy (MMD). MMD computes the average distance of the two samples mapped into the reproducing kernel Hilbert space (RKHS) of a universal kernel (e.g., Gaussian kernel). MMD critically relies on the choice of the feature representation (i.e., the kernel function) and thus might fail for complex, structured data such as sequences or images, and other data where deep learning excels.

Another non-parametric two-sample test is the classifier two-sample test (C2ST). C2ST splits the data into two chunks, training a classifier on one part and evaluating it on the remaining data. If the classifier predicts significantly better than chance, the test rejects the null hypothesis. Since a part of the data set needs to be put aside for training, not the full data set is used for computing the test statistic, which limits the power

*We provide code at `https://github.com/mkirchler/deep-2-sample-test`

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

of the method. Furthermore, the performance of the method depends on the selection of the train-test split.

In this work, we propose a two-sample testing procedure that uses deep learning to obtain a suitable data representation. It first maps the data onto a hidden-layer of a deep neural network that was trained (in an unsupervised or supervised fashion) on an independent, auxiliary data set, and then it performs a location test. Thus we are able to work on any kind of data that neural networks can work on, such as audio, images, videos, time-series, graphs, and natural language. We propose two test statistics that can be evaluated in linear time (in the number of observations), based on MMD and Fisher discriminant analysis, respectively. We derive asymptotic distributions of both test statistics. Our theoretical analysis proves that the two-sample test procedure asymptotically controls the type-1 error rate, has asymptotically vanishing type-2 error rate and is robust both with respect to transfer learning and approximate training.

We empirically evaluate the proposed methodology in a variety of applications from the domains of computational musicology, computer vision, and neuroimaging. In these experiments, the proposed deep two-sample tests consistently outperform the closest competing method (including deep kernel methods and C2STs) by up to 35 percentage points in terms of the type-2 error rate, while properly controlling the type-1 error rate.

2 PROBLEM STATEMENT & NOTATION

We consider non-parametric two-sample statistical testing, that is, to answer the question whether two samples are drawn from the same (unknown) distribution or not. We distinguish between the case that the two samples are drawn from the same distribution (the null hypothesis, denoted by H_0) and the case that the samples are drawn from different distributions (the alternative hypothesis H_1).

We differentiate between type-1 errors (i.e,rejecting the null hypothesis although it holds) and type-2 errors (i.e., not rejecting H_0 although it does not hold). We strive for both the type-1 error rate to be upper bounded by some significance level α, and the type-2 error rate to converge to 0 for unlimited data. The latter property is called consistency and means that with sufficient data, the test can reliably distinguish between any pair of probability distributions.

Let p, q, p' and q' be probability distributions on $\mathbb{R}^d$ with common dominating Borel measure μ. We abuse notation somewhat and denote the densities with respect to μ also by p, q, p' and q'. We want to perform a two-sample test on data drawn from p and q, i.e. we test the null hypothesis $H_0 : p = q$ against the alternative $H_1 : p \neq q$. p' and q' are assumed to be in some sense similar to p and q, respectively, and act as auxiliary task for tuning the test (the case of $p = p'$ and $q = q'$ is perfectly valid, in which case this is equivalent to a data splitting technique).

We have access to four (independent) sets $\mathcal{X}_n, \mathcal{Y}_n, \mathcal{X}'_{n'}$, and $\mathcal{Y}'_{n'}$ of observations drawn from p, q, p', and q', respectively. Here $\mathcal{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ and $X_i \sim p$ for all i (analogue definitions hold for $\mathcal{Y}_n, \mathcal{X}'_{n'}$, and $\mathcal{Y}'_{n'}$). Empirical averages with respect to a function f are denoted by $\overline{f(\mathcal{X}_n)} := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

We investigate function classes of deep ReLU networks with a final tanh activation function:

$$\mathcal{TF}_N := \Big\{ \tanh \circ W_{D-1} \circ \sigma \circ \ldots \circ \sigma \circ W_1 : \mathbb{R}^d \to \mathbb{R}^H \Big| \\ W_1 \in \mathbb{R}^{H \times d}, W_j \in \mathbb{R}^{H \times H} \text{ for } j = 2, \ldots, D-1, \\ \prod_{j=1}^{D-1} ||W_j||_{Fro} \leq \beta_N, D \leq D_N \Big\}$$

Here, the activation functions tanh and $\sigma(z) := \mathrm{ReLU}(z) = \max(0, z)$ are applied elementwise, $||\cdot||_{Fro}$ is the Frobenius norm, $H = d + 1$ is the width and D_N and β_N are depth and weight restrictions onto the networks. This can be understood as the mapping onto the last hidden layer of a neural network concatenated with a tanh activation.

3 DEEP TWO-SAMPLE TESTING

In this section, we propose two-sample testing based on two novel test statistics, the **Deep Maximum Mean Discrepancy (DMMD)** and the **Deep Fisher Discriminant Analysis (DFDA)**. The test asymptotically controls the type-1 error rate, and it is consistent (i.e., the type-2 error rate converges to 0). Furthermore, we will show that consistency is preserved under both transfer learning on a related task, as well as only approximately solving the training step.

3.1 Proposed Two-sample Test

Our proposed test consists of the following two steps. 1. We train a neural network over an auxiliary *training* data set. 2. We then evaluate the maximum mean discrepancy test statistic (Gretton et al., 2012a) (or a variant of it) using as kernel the mapping from the input domain onto the network's last hidden layer.

3.1.1 Training Step

Let the training data be $\mathcal{X}'_{n'}$ and $\mathcal{Y}'_{m'}$. Denote $N = n' + m'$. We run a (potentially inexact) training algorithm

to find $\phi_N \in \mathcal{TF}_N$ with:

$$\left\| \frac{1}{N} \left(\sum_{i=1}^{n'} \phi_N(X_i') - \sum_{i=1}^{m'} \phi_N(Y_i') \right) \right\| + \eta \geq \max_{\phi \in \mathcal{TF}_N} \left\| \frac{1}{N} \left(\sum_{i=1}^{n'} \phi(X_i') - \sum_{i=1}^{m'} \phi(Y_i') \right) \right\|.$$

Here, $\eta \geq 0$ is a fixed leniency parameter (independent of N); finding true global optima in neural networks is a hard problem, and an $\eta > 0$ allows us to settle with good-enough, local solutions. This procedure is also related to the early-stopping regularization technique, which is commonly used in training deep neural networks (Prechelt, 1998).

3.1.2 Test Statistic

We define the mean distance of the two test populations $\mathcal{X}_n, \mathcal{Y}_m$ measured on the hidden layer of a network ϕ as

$$D_{n,m}(\phi) := \overline{\phi(\mathcal{X}_n)} - \overline{\phi(\mathcal{Y}_m)}.$$

Using ϕ_N from the training step, we define the Deep Maximum Mean Discrepancy (DMMD) test statistic as

$$S_{n,m}(\phi_N, \mathcal{X}_n, \mathcal{Y}_m) := \frac{nm}{n+m} \left|\left| D_{n,m}(\phi_N) \right|\right|^2.$$

We can normalize this test statistic by the (inverse) empirical covariance matrix:

$$T_{n,m}(\phi_N, \mathcal{X}_n, \mathcal{Y}_m) := \frac{nm}{n+m} D_{n,m}(\phi_N)^\top \hat{\Sigma}_{n,m}^{-1} D_{n,m}(\phi_N).$$

This leads to a test statistic (which we call Deep Fisher Discriminant Analysis—DFDA) with an asymptotic distribution that is easier to evaluate. Note that the empirical covariance matrix is defined as:

$$\hat{\Sigma}_{n,m} := \hat{\Sigma}_{n,m}(\phi_N) := \frac{1}{n+m-1} \sum_{i=1}^{m+n} (\phi_N(Z_i) - \overline{\phi_N(\mathcal{Z})})(\phi_N(Z_i) - \overline{\phi_N(\mathcal{Z})})^\top + \rho_{n,m} I,$$

where $\rho_{n,m} > 0$ is a factor guaranteeing numerical stability and invertibility of the covariance matrix, and $\mathcal{Z} = \{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$.

3.1.3 Discussion

Intuitively, we map the data onto the last hidden layer of the neural network and perform a multivariate location test on whether both map to the same location. If the distance $D_{n,m}$ between the two means is too large, we reject the hypothesis that both samples are drawn from the same distribution. Consistency of this procedure is guaranteed by the training step.

Interpretation as Empirical Risk Minimization If we identify X_i' with $(Z_i', 1)$ and Y_i' with $(Z_{n'+i}', -1)$ in a regression setting, this is equivalent to an (inexact) empirical risk minimization with loss function $L(t, \hat{t}) = 1 - t\hat{t}$:

$$\max_\phi \left\| \frac{1}{N} \sum_{i=1}^{N} t_i' \phi(Z_i') \right\| = \max_\phi \max_{||w|| \leq 1} \frac{1}{N} \sum_{i=1}^{N} t_i' w^\top \phi(Z_i'),$$

which is equivalent to

$$\min_\phi \min_{||w|| \leq 1} R_N'(w^\top \phi) := \frac{1}{N} \sum_{i=1}^{N} L(t_i', w^\top \phi(Z_i')), \quad (1)$$

where we denote by R_N' the empirical risk; the corresponding expected risk is $R'(f) = \mathbb{E}[1 - t' f(Z')]$. Assuming that $\Pr(t' = 1) = \Pr(t' = -1) = \frac{1}{2}$, we have for the Bayes risk $R'^* = \inf_{f:\mathbb{R}^d \to [-1,1]} R'(f) = 1 - \epsilon'$ with $\epsilon' > 0$ if and only if $p' \neq q'$. As long as p' and q' are selected close enough to p and q, respectively, the corresponding test will be able to distinguish between the two distributions.

Since we discard w after optimization and use the norm of the hidden layer on the test set again, this implies some fine-tuning on the test data, without compromising the test statistic (see Theorem 3.1 below). This property is especially helpful in neural networks, since for practical transfer learning, only fine-tuning the last layer can be extremely efficient, even if the transfer and actual task are relatively different (Lu et al., 2015).

Relation to kernel-based tests The test statistic $S_{n,m}$ is a special case of the standard squared Maximum Mean Discrepancy (Gretton et al., 2012b) with the kernel $k(z_1, z_2) := \langle \phi(z_1), \phi(z_2) \rangle$ (analogously for $T_{n,m}$ and the Kernel FDA Test (Harchaoui et al., 2008)). For a fixed feature map ϕ this kernel is not characteristic, and hence the resulting test not necessarily consistent for arbitrary distributions p, q. However, by first choosing ϕ in a data-dependent way, we can still achieve consistency.

3.2 Control of Type-1 Error

Due to our choice of ϕ_N, there need not be a unique, well-defined limiting distribution for the test statistics when $n, m \to \infty$. Instead, we will show that for each *fixed* ϕ, the test statistic $S_{n,m}$ has a well-defined limiting distribution that can be well evaluated. If in addition the covariance matrix is invertible, then the same holds for $T_{n,m}$.

In particular, the following theorem will show that $D_{n,m}(\phi)$ converges towards a multivariate normal distribution for $n, m \to \infty$. $S_{n,m}$ then is asymptotically

distributed like a weighted sum of χ^2 variables, and $T_{n,m}$ like a χ^2_H (again, if well-defined).

Theorem 3.1. *Let $p = q$, $\phi \in \mathcal{TF}$ and $\Sigma := Cov(\phi(X_1))$ and assume that $\frac{n}{n+m} \to r \in (0,1)$ as $n, m \to \infty$.*

(i) As $n, m \to \infty$, it holds that

$$\sqrt{\frac{mn}{m+n}} D_{n,m}(\phi) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

(ii) As $n, m \to \infty$,

$$S_{n,m}(\phi, \mathcal{X}_n, \mathcal{Y}_m) \xrightarrow{d} \sum_{i=1}^{H} \lambda_i \xi_i^2,$$

where $\xi_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and λ_i are the eigenvalues of Σ.

(iii) If additionally Σ is invertible, and $\rho_{n,m} \downarrow 0$ then as $n, m \to \infty$

$$T_{n,m}(\phi, \mathcal{X}_n, \mathcal{Y}_m) \xrightarrow{d} \chi^2_H.$$

Sketch of proof (full proof in Appendix A.1). (i) As under H_0 $\phi(X_i)$ and $\phi(Y_j)$ are identically distributed, $D_{n,m}(\phi)$ is centered and one can show the result using a Central Limit Theorem.

(ii) and *(iii)* then follow from the continuous mapping theorem and properties of the multivariate normal distribution. □

Under some additional assumptions we can also use a Berry-Esseen type of result to quantify the quality of the normal approximation of $D_{n,m}(\phi_N)$ conditioned on the training. In particular, if we assume that $n = m$ and $\Sigma = \mathrm{Cov}_{p,q}(\phi_N(X_1)) | \mathcal{X}'_n, \mathcal{Y}'_n$ invertible, then Bentkus (2005) shows that the normal approximation on convex sets is $\mathcal{O}\left(\frac{H^{1/4}}{\sqrt{n}}\right)$. Computing p-values for both $S_{n,n}$ and $T_{n,n}$ only requires computation over convex sets, so the result is directly applicable.

3.2.1 Computational Aspects

Testing with $S_{n,m}$ As shown in Theorem 3.1, the null distribution of $S_{n,m}$ can be approximated as the weighted sum of independent χ^2-variables. There are several approaches to computing the cumulative distribution function of this distribution, see Bausch (2013) for an overview and Zhou and Guan (2018) for an implementation. However, computing p-values with this method can be rather costly.

Alternatively, note that the test statistic $S_{n,m}$ is linear in the number of observations and dimensions. Hence, estimating the null distribution via Monte-Carlo permutation sampling (Ernst et al., 2004) is feasible. Note also that it suffices to evaluate the feature map ϕ on each data point only once and then permute the class labels, saving more time.

In practice we found that the resampling-based test performed considerably faster. Hence, in the remainder of this work, we will evaluate the null hypothesis of the DMMD via the resampling method.

Testing with $T_{n,m}$ Since in many practical situations one wants to use standard neural network architectures (such as ResNets), the number of neurons in the last hidden layer H may be rather large, compared to n, m. Therefore, using the full, high-dimensional hidden layer representation might lead to suboptimal normal approximations. Instead, we propose to use a principal component analysis on the feature representation $(\phi(Z_i))_{i=1}^{n+m}$ to reduce the dimensionality to $\hat{H} \ll m + n$. In fact, this does not break the asymptotic theory derived in Theorem 3.1, even though the PCA is both trained and evaluated on the test data; details can be found in Appendix C. Unfortunately, the $\mathcal{O}\left(\frac{H^{1/4}}{\sqrt{n}}\right)$ rate of convergence is not valid anymore, due to the observations not being independent. We still need to grow $\hat{H}$ towards H with n, m in order for the consistency results in the next section to hold, however. Empirically we found $\hat{H} = \min\left(\sqrt{\frac{n+m}{2}}, H\right)$ to perform well.

The cumulative distribution function of the χ^2_H distribution can be evaluated very efficiently. Although for the DFDA it is also possible to estimate the null hypothesis via a Monte Carlo permutation scheme, doing so is more costly than for the DMMD, since it involves either a matrix inversion once or solving a linear system for each permutation draw. Hence, in this work we focus on using the asymptotic distribution.

3.3 Consistency

In this section we show that if (a), the restrictions β_N, D_N on weights and depth of networks in $\mathcal{TF}_N$ are carefully chosen, (b), the transfer task is not too far from the original task, and (c), the leniency parameter η in the training step is small enough, then our proposed test is consistent, meaning the type-2 error rate converges to 0.

Theorem 3.2. *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 $\epsilon' > 0$, and assume that the following holds:*

(i) $\frac{\beta_N^2 D_N}{N} \to 0$, $\beta_N \to \infty$ and $D_N \to \infty$ for $N \to \infty$ for the parameters of the function classes $\mathcal{TF}_N$,

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

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

(iv) p' *and* q' *have bounded support on* $\mathbb{R}^d$.†

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

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

Sketch of proof (full proof in Appendix A.2). The test statistics $S_{n,m}$ is lower-bounded by a rescaled version of $\sqrt{N}(1 - R_{n,m}(\psi_N))$, where $\psi_N = w_N^\top \phi_N$ with w_N selected as in (1). Then, if $1 - R_{n,m}(\psi_N) \geq c > 0$, the test statistic diverges.

The finite-sample error $R_{n,m}(\psi_N)$ approaches its population version $R(\psi_N)$ for large n, m, and the difference between $R(\psi_N)$ and $R'(\psi_N)$ can be controlled over δ. The rest of the proof is akin to standard consistency proofs in regression and classification. Namely, we can split $R'_N(\psi_N) - R'^*$ into approximation and estimation error and control these via a Universal Approximation Theorem (Hanin, 2017), and Rademacher complexity bounds on the neural network function class (Golowich et al., 2017), respectively. □

The main caveat of Theorem 3.2 is that it gives no explicit directions to choose the transfer task p' and q'. Whether the respective μ-densities are L_1-close to the testing densities in general cannot be answered, and similarly the Bayes error rate $1 - \epsilon'$ is not known beforehand. If abundant data for the testing task is at hand, then splitting the data is the safe way to go; if data is scarce, Theorem 3.2 gives justification that a *reasonably close* transfer task will have good power as well.

The bounded support requirement *(iv)* on p' and q' can be circumvented as well – by choosing the support large enough one can always just truncate (X'_i) and (Y'_i) and will still satisfy requirements *(ii)* and *(iii)*, especially also in the case of $p' = p$ and $q' = q$ with unbounded support. This procedure, however, requires knowledge of where to truncate the transfer distributions. Instead one can also grow the support of p' and q' with N; for more details, see Appendix B.

†A similar Theorem holds also for the case of unbounded support, see Appendix B

4 RELATED WORK

In this section, we give an overview over the state-of-the-art in non-parametric two-sample testing for high-dimensional data.

Kernel Methods The methods most related to our method are the kernelized maximum mean discrepancy (MMD) (Gretton et al., 2012a) and the kernel Fisher discriminant analysis (KFDA) (Harchaoui et al., 2008). Both methods effectively metricize the space of probability distributions by mapping distribution features onto mean embeddings in universal reproducing kernel Hilbert spaces (RKHS, (Steinwart and Christmann, 2008)). Test statistics derived from these mean embeddings can be efficiently evaluated using the kernel trick (in quadratic time in the number of observations, although there are lower-powered linear-time variations). Mean Embeddings (ME) and Smoothed Characteristic Functions (SCF) (Chwialkowski et al., 2015; Jitkrittum et al., 2016) are kernel-based linear-time test statistics that are (almost surely) proper metrics on the space of probability distributions. All four methods rely on characteristic kernels to yield consistent tests and are closely related.

Deep Kernel Methods In the context of training and evaluating Generative Adversarial Networks (GANs), several authors have investigated the use of the MMD with kernels parametrized by deep neural networks. In Bińkowski et al. (2018); Li et al. (2017); Arbel et al. (2018), the authors feed features extracted from deep neural networks into characteristic kernels. Jitkrittum et al. (2018) use deep kernels in the context of relative goodness-of-fit testing without directly considering consistency aspects of this approach. Extensions from the GAN literature to two-sample testing is not straightforward since statistical consistency guarantees strongly depend on careful selection of the respective function classes. To the best of our knowledge, all previous works made simplifying assumptions on injectivity or even invertibility of the involved networks.

In this work we show that a linear kernel on top of transfer-learned neural network feature maps (as has also been done by Xu et al. (2018) for GAN evaluation) is not only sufficient for consistency of the test, but also performs considerably better empirically in all settings we analyzed. In addition to that, our test statistics can be directly evaluated in linear instead of quadratic time (in the sample size) and the corresponding asymptotic null distributions can be exactly computed (in contrast to the MMD & KFDA).

Classifier Two-Sample Tests (C2ST) First proposed by Friedman (2003) and then further analyzed by

Kim et al. (2016) and Lopez-Paz and Oquab (2016), the idea of the C2ST is to utilize a generic classifier, such as a neural network or a k-nearest neighbor approach for the two-sample testing problem. In particular, they split the available data into training and test set, train a classifier on the training set and evaluate whether the performance on the test set exceeds random variation. The main drawback of this approach is that the data has to be split in two chunks, creating a trade-off: if the training set is too small, the classifier is unlikely to find a statistically relevant signal in the data; if the training set is large and thus the test set small, the C2ST test loses power.

Our method circumvents the need to split the data in training and test set – Theorem 3.2 shows that training on a reasonably close transfer data set is sufficient. Even more, as shown in Section 3.1.3, our method can be interpreted as empirical risk minimization with additional fine-tuning of the last layer on the testing data, guaranteed to be as least as good as an equivalent method with fixed last layer.

5 EXPERIMENTS

In this section, we compare our proposed deep learning two-sample tests with other state-of-the-art approaches.

5.1 Experimental setup

For the **DFDA** and **DMMD** tests we train a deep neural network on a related task; details will be deferred to the corresponding sections. We report both the performance of the deep MMD $S_{n,m}$ where we estimate the null hypothesis via a Monte Carlo permutation sample (Ernst et al., 2004) (we fix $M = 1000$ resampling permutations except otherwise noted), and the deep FDA statistic $T_{n,m}$, for which we use the asymptotic χ^2_H distribution. As explained in Section 3.2.1, for the DFDA we project the last hidden layer onto $\hat{H} < H$ dimensions using a PCA. We found the heuristic $\hat{H} := \sqrt{\frac{m+n}{2}}$ to perform well across a number of tasks (disjoint from the ones presented in this section). For the DMMD we do not need any dimensionality reduction. We calibrated parameters of both tests on data disjoint from the ones that we report results on in the subsequent sections.

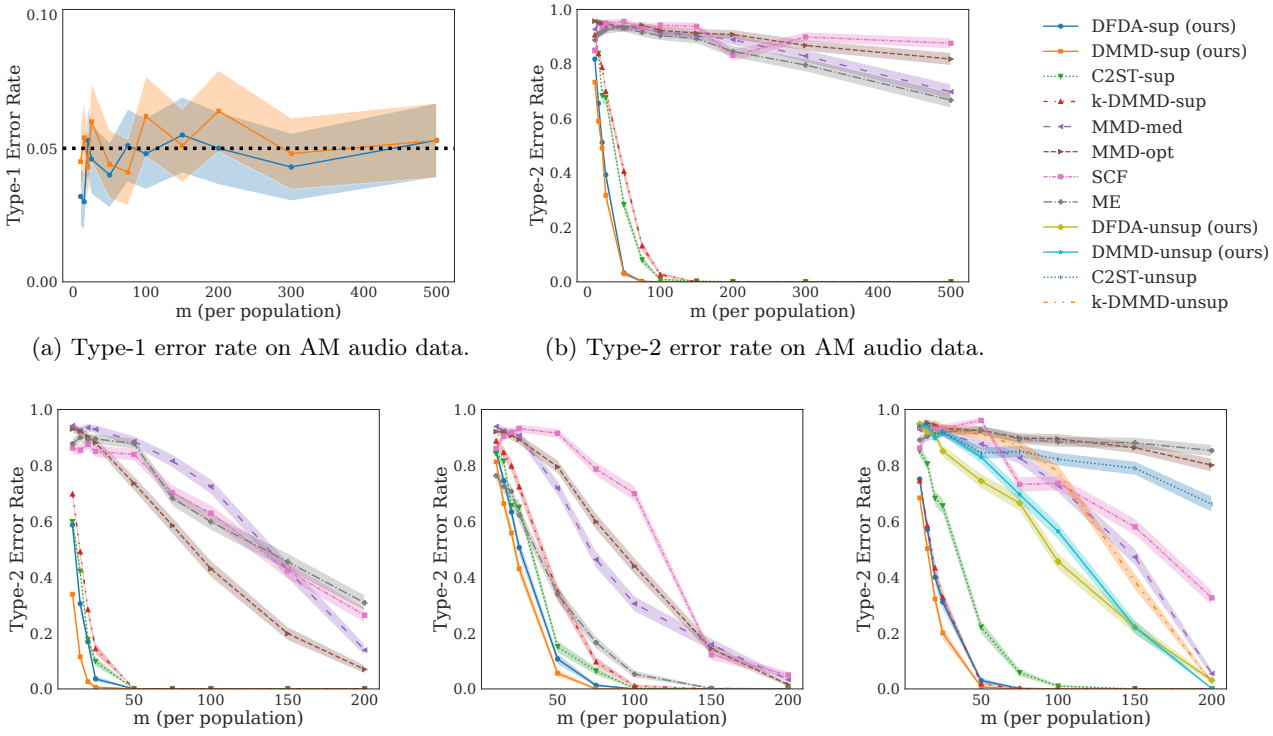
For the **C2ST**, we train a standard logistic regression on top of the pretrained features extracted from the same neural network as for our methods.

For the **kernel MMD** we report two kernel bandwidth selection strategies for the Gaussian kernel. The first variant is the "median distance" heuristic (Gretton et al., 2012a) which selects the median of the euclidean distances of all data points (MMD-med). The second variant, reported by Gretton et al. (2012b), splits the data in two disjoint sets and selects the bandwidth that maximizes power on the first set and evaluates the MMD on the second set (MMD-opt). We use the implementation provided by Jitkrittum et al. (2016), which estimates the null hypothesis via a Monte Carlo permutation scheme (we again use $M = 1000$ permutations).

For the **Smoothed Characteristic Functions** (SCF) and **Mean Embeddings** (ME), we select the number of test locations based on the task and sample size. The locations are selected either randomly (as presented by Chwialkowski et al. (2015)) or optimized on half of the data via the procedure described by Jitkrittum et al. (2016). The kernel was either selected using the median heuristic, or via a grid search as by Chwialkowski et al. (2015); Jitkrittum et al. (2016). In each case we report the kernel and location selection method that performed best on the given task, with details given in the corresponding paragraphs. Note that for very small sample sizes, both SCF and ME oftentimes do not control the type-1 error rate properly, since they were designed for larger sample sizes. This results in highly variable type-2 error rate for small m in the experiments. Again, we use the implementation provided by Jitkrittum et al. (2016).

In addition to these published methods, we also compare our method against a **deep kernel MMD test** (k-DMMD), i.e. the MMD test where the output of a pretrained neural network gets fed into a Gaussian kernel (instead of a linear kernel as in our case). Jitkrittum et al. (2018) used this method for relative goodness-of-fit testing instead of two-sample testing. For image data, we select the bandwidth parameter for the Gaussian kernel via the median heuristic, and for audio data via the power maximization technique (in each case the other variant performs considerably worse); the pretrained networks are the same as for our tests and the C2ST.

All experiments were run over 1000 runs. Type-1 error rates are estimated by drawing both samples (without replacement) from the same class and computing the rate of rejections. Similarly, type-2 error rates are estimated as the rate of not rejecting the null hypothesis when sampling from two distinct classes. All figures of type-1 and type-2 error rates show the 95% confidence interval based on a Wilson Score interval (and a "rule-of-three" approximation in the case of 0-values (Eypasch et al., 1995)). In all settings we fixed the significance level at $\alpha = 0.05$. In addition to that we show in Appendix D.3 empirically that also for smaller significance levels high power can be preserved. Preprocessing for image data is explained in Appendix D.2.



(a) Type-1 error rate on AM audio data.

(b) Type-2 error rate on AM audio data.

(c) Type-2 error rate on aircraft data.

(d) Type-2 error rate on KDEF data.

(e) Type-2 error rate on dogs data.

Figure 1: Results on AM audio (top row) and natural image (bottom row) data sets. Suffixes "-sup" indicate supervised pretraining, "-unsup" indicates unsupervised pretraining.

5.2 Control of Type-1 Error Rate

Since the presented test procedures are not exact tests it is important to verify that the type-1 error rate is controlled at the proper level. Figure 1a shows that the empirical type-1 error rate is well controlled for the amplitude modulated audio data introduced in the next section. For the other data sets, results are provided in Appendix D.4.

5.3 Power Analysis

Amplitude Modulated Audio Data Here we analyze the proposed test on the amplitude modulated audio example from (Gretton et al., 2012b). The task in this setting is to distinguish snippets from two different songs after they have been amplitude modulated (AM) and mixed with noise. We use the same preprocessing and amplitude modulation as Gretton et al. (2012b). We use the freely available music from Gramatik (2014); distribution p is sampled from track four, distribution q from track five and the remaining tracks on the album were used for training the network in a multi-class classification setting. As our neural network architecture we use a simple convolutional network, a variant from Dai et al. (2017), called M5 therein; see Appendix D.6 for details.

Figure 1b reports the results with varying number of observations under constant noise level $\sigma^2 = 1$. Our method shows high power, even at low sample sizes, whereas kernel methods need large amounts of data to deal with the task. Note that these results are consistent with the original results in Gretton et al. (2012b), where the authors fixed the sample size at $m = 10,000$ and consequently only used the (significantly less powerful) linear-time MMD test.

Aircraft We investigate the Fine-Grained Visual Classification of Aircraft data set (Maji et al., 2013). We select two visually similar aircraft families, namely Boeing 737 and Boeing 747 as populations p and q, respectively. The neural network embeddings are extracted from a ResNet-152 (He et al., 2016) trained on ILSVRC (Russakovsky et al., 2015). Figure 1c shows that all neural network architectures perform considerably better than the kernel methods. Furthermore, our proposed tests can also outperform both the C2ST and the deep kernel MMD.

Facial Expressions The Karolinska Directed Emotional Faces (KDEF) data set (Lundqvist et al., 1998)

Table 1: Results on neuroimaging data, comparing subjects who are cognitive normal (CN), have mild cognitive impairment (MCI) or have Alzheimer's disease (AD). *APOE* has neutral variant $\varepsilon 3$ and risk-factor variant $\varepsilon 4$. Numbers in parentheses denote sample size.

X (# obs)	Y (# obs)	p-value
CN (490)	AD (314)	$9.49 \cdot 10^{-5}$
CN (490)	MCI (287)	$2.44 \cdot 10^{-4}$
MCI (287)	AD (314)	$1.45 \cdot 10^{-3}$
APOE $\varepsilon 3$ (811)	*APOE* $\varepsilon 4$ (152)	$1.40 \cdot 10^{-2}$

has been previously used by Jitkrittum et al. (2016); Lopez-Paz and Oquab (2016). The task is to distinguish between faces showing positive (happy, neutral, surprised) and negative (afraid, angry, disgusted) emotions. The feature embeddings are again obtained from a ResNet-152 trained on ILSVRC. Results can be found in Figure 1d. Even though the images in ImageNet and KDEF are very different, the neural network tests again outperform the kernel methods. Also note that the apparent advantage of the mean embedding test for low sample sizes is due to an unreasonably high type-1 error rate (> 0.11 and > 0.085 at $m = 10, 15$, respectively).

Stanford Dogs Lastly, we evaluate our tests on the Stanford Dogs data set (Khosla et al., 2011), consisting of 120 classes of different dog breeds. As test classes we select the dog breeds 'Irish wolfhound' and 'Scottish deerhound', two breeds that are visually extremely similar. Since the data set is a subset of the ILSVRC data, we cannot train the networks on the whole ImageNet data again. Instead, we train a small 6-layer convolutional neural network on the remaining 118 classes in a multi-class classification setting and use the embedding from the last hidden layer. To show that our tests can also work with unsupervised transfer-learning, we also train a convolutional autoencoder on this data; the encoder part is identical to the supervised CNN, see Appendix D.7 for details. Note that for this setting, the theoretical consistency guarantees from Theorem 3.2 do not hold, although the type-1 error rate is still asymptotically controlled. Figure 1e reports the results, with *-sup denoting the supervised, and *-unsup the unsupervised transfer-learning task. As expected, tests based on the supervised embedding approach outperform other tests by a large margin. However, the unsupervised DMMD and DFDA still outperform kernel-based tests. Interestingly, both the C2ST and the k-DMMD method seem to suffer more severely from the mediocre feature embedding than our tests. One potential explanation for this phenomenon is the ability of DMMD and DFDA to fine-tune on the test data without the need to perform a data split.

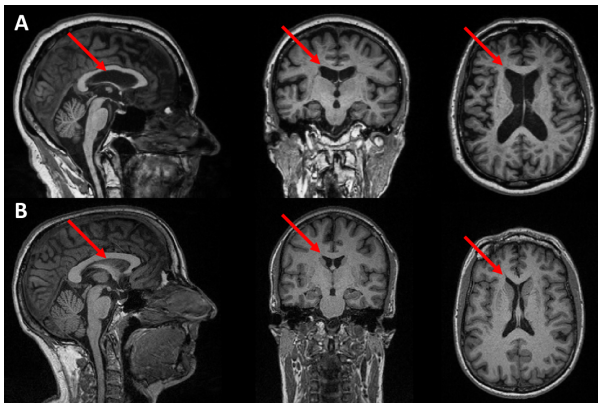


Figure 2: Slices of 3D-MRI scans of an Alzheimer's disease patient (A) and a cognitively normal individual (B). Note the enlargement of the lateral ventricles (indicated by red arrows) in the Alzheimer's disease patient.

Three-dimensional Neuroimaging Data In this section, we apply the DFDA test procedure to 3D Magnetic Resonance Imaging (MRI) scans and genetic information from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005). To this end, we transfer a 3D convolutional autoencoder that has been trained on MRI scans from the Brain Genomics Superstruct Project (Holmes et al., 2015) to perform statistical testing on the ADNI data. Details on preprocessing and network architecture are provided in Appendix D.10.

The ADNI dataset consists of individuals diagnosed with Alzheimer's Disease (AD), with Mild Cognitive Impairment (MCI), or as cognitively normal (CN); Figure 2 shows exemplaric images of an AD and a CN subject. Table 1 shows that our test can detect statistically significant differences between MRI scans of individuals with a different diagnosis. Additionally, we evaluate whether our test can detect differences between individuals who have a known genetic risk factor for neurodegenerative diseases and individuals without that risk factor. In particular, we compare the two variants $\varepsilon 3$ (the "normal" variant) and $\varepsilon 4$ (the risk-factor variant) in the Apolipoprotein E (*APOE*) gene, which is related to AD and other diseases (Corder et al., 1993). By grouping subjects according to which variant they exhibit we test for statistical dependence between a (binary) genetic mutation and (continuous) variation in 3D MRI scans. Table 1 shows that individuals with $\varepsilon 4$ and $\varepsilon 3$ *APOE* variants are significantly different, suggesting a statistical dependence between genetic variation and structural brain features.

Acknowledgements

The authors thank Stefan Konigorski and Jesper Lund for helpful discussions and comments. Marius Kloft acknowledges support by the German Research Foundation (DFG) award KL 2698/2-1 and by the Federal Ministry of Science and Education (BMBF) awards 031L0023A, 01IS18051A, and 031B0770E. Part of the work was done while Marius Kloft was a sabbatical visitor of the DASH Center at the University of Southern California. This work has been funded by the Federal Ministry of Education and Research (BMBF, Germany) in the project KI-LAB-ITSE (project number 01|S19066).

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database `adni.loni.usc.edu`. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Data collection and sharing of ADNI was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica Inc; Biogen Idec Inc; Bristol-Myers Squibb Company; Eisai Inc; Elan Pharmaceuticals Inc; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech Inc; GE Healthcare; Innogenetics N.V.; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace Inc; Merck & Co Inc; Meso Scale Diagnostics LLC; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc; Piramal Imaging; Servier; Synarc Inc; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`http://www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Samples from the National Cell Repository for AD (NCRAD), which receives government support under a cooperative agreement grant (U24 AG21886) awarded by the National Institute on Aging (AIG), were used in this study. Funding for the WGS was provided by the Alzheimer's Association and the Brin Wojcicki Foundation.

References

Michael Arbel, Dougal Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans. In *Advances in Neural Information Processing Systems*, pages 6700–6710, 2018.

Johannes Bausch. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *Journal of Physics A: Mathematical and Theoretical*, 46(50): 505202, 2013.

Vidmantas Bentkus. A lyapunov-type bound in rd. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

EH Corder, AM Saunders, WJ Strittmatter, DE Schmechel, PC Gaskell, GW Small, AD Roses, JL Haines, and MA Pericak-Vance. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5):921–923, 1993.

Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425. IEEE, 2017.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Michael D Ernst et al. Permutation methods: a basis for exact inference. *Statistical Science*, 19(4):676–685, 2004.

Ernst Eypasch, Rolf Lefering, CK Kum, and Hans Troidl. Probability of adverse events that have not yet occurred: a statistical reminder. *Bmj*, 311(7005): 619–620, 1995.

Jerome Friedman. On multivariate goodness-of-fit and two-sample testing. In *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, page 311, 2003.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Gramatik. The age of reason. `http://dl.lowtempmusic.com/Gramatik-TAOR.zip`, 2014. [Online; accessed May/23/2019].

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

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012b.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *arXiv preprint arXiv:1708.02691*, 2017.

Zaïd Harchaoui, Francis R Bach, and Èric Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Avram J Holmes, Marisa O Hollinshead, Timothy M O'Keefe, Victor I Petrov, Gabriele R Fariello, Lawrence L Wald, Bruce Fischl, Bruce R Rosen, Ross W Mair, Joshua L Roffman, et al. Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific data*, 2:150031, 2015.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In *Advances in Neural Information Processing Systems*, pages 808–819, 2018.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, 80:14–23, 2015.

Daniel Lundqvist, Anders Flykt, and Arne Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Bettina Mieth, Marius Kloft, Juan Antonio Rodríguez, Sören Sonnenburg, Robin Vobruba, Carlos Morcillo-Suárez, Xavier Farré, Urko M Marigorta, Ernst Fehr, Thorsten Dickhaus, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific reports*, 6:36671, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005.

J Neyman and ES Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical*

Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231:289–337, 1933.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

Hao Zhou, Vamsi K Ithapu, Sathya Narayanan Ravi, Vikas Singh, Grace Wahba, and Sterling C Johnson. Hypothesis testing in unsupervised domain adaptation with applications in alzheimer's disease. In *Advances in neural information processing systems*, pages 2496–2504, 2016.

Quan Zhou and Yongtao Guan. On the null distribution of bayes factors in linear regression. *Journal of the American Statistical Association*, 113(523): 1362–1371, 2018.