MAGDiff: Covariate Data Set Shift Detection via Activation Graphs of Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Despite their successful application to a variety of tasks, neural networks remain limited, like other machine learning methods, by their sensitivity to shifts in the data: their performance can be severely impacted by differences in distribution between the data on which they were trained and that on which they are deployed. In this article, we propose a new family of representations, called MAGDiff, that we extract from any given neural network classifier and that allows for efficient covariate data shift detection without the need to train a new model dedicated to this task. These representations are computed by comparing the activation graphs of the neural network for samples belonging to the training distribution and to the target distribution, and yield powerful data- and task-adapted statistics for the two-sample tests commonly used for data set shift detection. We demonstrate this empirically by measuring the statistical powers of two-sample Kolmogorov-Smirnov (KS) tests on several different data sets and shift types, and showing that our novel representations induce significant improvements over a state-of-the-art baseline relying on the network output.

1 Introduction

During the last decade, neural networks (NN) have become immensely popular, reaching state-of-the-art performances in a wide range of situations. Nonetheless, once deployed in real-life settings, NN can face various challenges such as being subject to adversarial attacks (Huang et al., 2017), being exposed to out-of-distributions samples (samples that were not presented at training time) (Hendrycks & Gimpel, 2016), or more generally being exposed to a *distribution shift*: when the distribution of inputs differs from the training distribution (*e.g.*, input objects are exposed to a corruption due to deterioration of measure instruments such as cameras or sensors). Such distribution shifts are likely to degrade performances of presumably well-trained models (Wiles et al., 2021b), and being able to detect such shifts is a key challenge in monitoring NN once deployed in real-life applications. Though shift detection for univariate variables is a well-studied problem, the task gets considerably harder with high-dimensional data, and seemingly reasonable methods often end up performing poorly (Ramdas et al., 2014).

In this work, we introduce the Mean Activation Graph Difference (MAGDiff), a new approach that harnesses the powerful dimensionality reduction capacity of deep neural networks in a data- and task-adapted way. The key idea, further detailed in Section 4, is to consider the activation graphs generated by inputs as they are processed by a neural network that has already been trained for a classification task, and to compare such graphs to those associated to samples from the training distribution. The method can thus be straightforwardly added as a diagnostic tool on top of preexisting classifiers without requiring any further training ; it is easy to implement, and computationally inexpensive. As the activation graphs depend on the network weights, which in turn have been trained for the data and task at hand, one can also hope for them to capture information that is most relevant to the context. Hence, our method can easily support, and benefit from, any improvements in deep learning.

Our approach is to be compared to *Black box shift detection* (BBSD), a method introduced in Lipton et al. (2018); Rabanser et al. (2019) that shares a similar philosophy. BBSD uses the output of a trained classifier to efficiently detect various types of shifts (see also Section 4); in their experiments, BBSD generally beats other

methods, the runner-up being a much more complex and computationally costly multivariate two-sample test combining an autoencoder and the Maximum Mean Discrepancy statistic (Gretton et al., 2012).

Our contributions are summarized as follows.

1. Given any neural network classifier, we introduce a new family of representations `MAGDiff`, that is obtained by comparing the activation graphs of samples to the mean activation graph of each class in the training set.

2. We propose to use `MAGDiff` as a statistic for data set shift detection. More precisely, we combine our representations with the statistical method that was proposed and applied to the Confidence Vectors (CV) of classifiers in Lipton et al. (2018), yielding a new method for shift detection.

3. We experimentally show that our shift detection method with `MAGDiff` outperforms the state-of-the-art BBSD with CV on a variety of datasets, covariate shift types and shift intensities, often by a wide margin. Our code is provided in the Supplementary Material and will be released publicly.

2 Related Work

Detecting changes or outliers in data can be approached from the angle of anomaly detection, a well-studied problem (Chandola et al., 2009), or out-of-distribution (OOD) sample detection (Shafaei et al., 2018). Among techniques that directly frame the problem as shift detection, kernel-based methods such as Maximum Mean Discrepancy (MMD) (Gretton et al., 2012; Zaremba et al., 2013) and Kernel Mean Matching (KMM) (Gretton et al., 2009; Zhang et al., 2013) have proved popular, though they scale poorly with the dimensionality of the data (Ramdas et al., 2014). Using classifiers to test whether samples coming from two distributions can be correctly labeled, hence whether the distributions can be distinguished, has also been attempted; see, *e.g.*, Kim et al. (2021). The specific cases of covariate shift (Jang et al., 2022; Uehara et al., 2020; Rabanser et al., 2019) and label shift (Storkey, 2009; Lipton et al., 2018) have been further investigated, from the point of view of causality and anticausality (Schölkopf et al., 2012). Moreover, earlier investigations of similar questions have arisen from the fields of economics (Heckman, 1977) and epidemiology (Saerens et al., 2002).

Among the works cited above, Lipton et al. (2018) and Rabanser et al. (2019) are of particular interest to us. In Lipton et al. (2018), the authors detect label shifts using shifts in the distribution of the outputs of a well-trained classifier; they call this method Black Box Shift Detection (BBSD). In Rabanser et al. (2019), the authors observe that BBSD tends to generalize very well to covariate shifts, though without the theoretical guarantees it enjoys in the label shift case. Our method is partially related to BBSD. Roughly summarized, we apply similar statistical tests—combined univariate Kolmogorov-Smirnov tests—to different features—Confidence Vectors (CV) in the case of BBSD, distances to mean activation graphs (`MAGDiff`) in ours. Similar statistical ideas have also been explored in Alberge et al. (2019) and Bar-Shalom et al. (2022), while neural network activation graph features have been studied in, *e.g.*, Lacombe et al. (2021) and Horta et al. (2021). The related issue of the robustness of various algorithms to diverse types of shifts has been recently investigated in Wiles et al. (2021a).

3 Background

3.1 Shift Detection with Two-Sample Tests

There can often be a shift between the distribution $\mathbb{P}_0$ of data on which a model has been trained and tested and the distribution $\mathbb{P}_1$ of the data on which it is used after deployment; many factors can cause such a shift, *e.g.*, a change in the environment, in the data acquisition process, or the training set being unrepresentative. Detecting shifts is crucial to understanding, and possibly correcting, potential losses in performance; even shifts that do not strongly impact accuracy can be important symptoms of inaccurate assumptions or changes in deployment conditions.

Additional assumptions can sometimes be made on the nature of the shift. In the context of a classification task, where data points are of the shape (x, y) with x the feature vector and y the label, a shift that preserves the conditional distribution $p(x|y)$ (but allows the proportion of each label to vary) is called *label shift*. Conversely, a *covariate shift* occurs when $p(y|x)$ is preserved, but the distribution of $p(x)$ is allowed to change. In this article, we focus on the arguably harder case of covariate shifts. See Section 5 for examples of such shifts in numerical experiments.

Shifts can be detected using *two-sample tests*: that is, a statistical test that aims at deciding between the two hypotheses

$$H_0 : \mathbb{P}_0 = \mathbb{P}_1 \text{ and } H_1 : \mathbb{P}_0 \neq \mathbb{P}_1,$$

given two random sets of samples, X_0 and X_1, independently drawn from two distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ (see, *e.g.*, Heumann & Schomaker (2023) for an introduction to hypothesis testing). To do so, many statistics have been derived, depending on the assumptions made on $\mathbb{P}_0$ and $\mathbb{P}_1$. In the case of distributions supported on $\mathbb{R}$, one such test is the *univariate Kolmogorov-Smirnov (KS) test*, of which we make use in this article. Given, as above, two sets of samples $X_0, X_1 \subset \mathbb{R}$, consider the empirical distribution functions $F_i(z) := \frac{1}{\mathrm{Card}(X_i)} \sum_{x \in X_i} 1_{x \leq z}$ for $i = 0, 1$ and $z \in \mathbb{R}$, where Card denotes the cardinality. Then the statistic associated with the KS test and the samples is $T := \sup_{z \in \mathbb{R}} |F_0(z) - F_1(z)|$. If $\mathbb{P}_0 = \mathbb{P}_1$, the distribution of T is independent of $\mathbb{P}_0$ and converges to a known distribution when the sizes of the samples tend to infinity (under mild assumptions) (Smirnov, 1939). Hence approximate p-values can be derived. The KS test can also be used to compare multivariate distributions: if $\mathbb{P}_0$ and $\mathbb{P}_1$ are distributions on $\mathbb{R}^D$, a p-value p_i can be computed from the samples by comparing the i-th entries of the vectors of $X_0, X_1 \subset \mathbb{R}^D$ using the univariate KS test, for $i = 1, \ldots, D$. A standard and conservative way of combining those p-values is to reject H_0 if $\min(p_1, \ldots, p_D) < \alpha/D$, where α is the significance level of the test. This is known as the *Bonferroni correction* (Voss & George, 1995). Other tests tackle the multidimensionality of the problem more directly, such as the *Maximum Mean Discrepancy (MMD) test*, though not necessarily with greater success (see, *e.g.*, Ramdas et al. (2014)).

3.2 Neural Networks

We now recall the basics of *neural networks* (NN), which will be our main object of study. We define a neural network[1] as a (finite) sequence of functions called *layers* $f_1, \ldots, f_L$ of the form $f_\ell \colon \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_{\ell+1}}, x \mapsto \sigma_\ell(W_\ell \cdot x + b_\ell)$, where the parameters $W_\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ and $b_\ell \in \mathbb{R}^{n_{\ell+1}}$ are called the weight matrix and the bias vector respectively, and σ_ℓ is an (element-wise) activation map (*e.g.*, sigmoid or ReLU). The neural network encodes a map $F \colon \mathbb{R}^d \to \mathbb{R}^D$ given by $F = f_L \circ \cdots \circ f_1$. We sometimes use F to refer to the neural network as a whole, though it has more structure.

When the neural network is used as a classifier, the last activation function σ_L is often taken to be the *softmax* function, so that $F(x)_i$ can be interpreted as the confidence that the network has in x belonging to the i-th class, for $i = 1, \ldots, D$. For this reason, we use the terminology *confidence vector* (CV) for the output $F(x) \in \mathbb{R}^D$. The true class of x is represented by a label $y = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^D$ that takes value 1 at the coordinate indicating the correct class and 0 elsewhere. The parameters of each layer (W_ℓ, b_ℓ) are typically learned from a collection of training observations and labels $\{(x_n, y_n)\}_{n=1}^N$ by minimizing a cross-entropy loss through gradient descent, in order to make $F(x_n)$ as close to y_n as possible on average over the training set. The *prediction* of the network on a new observation x is then given by $\arg\max_{i=1,\ldots,D} F(x)_i$, and its (test) *accuracy* is the proportion of correct predictions on a new set of observations $\{(x'_n, y'_n)\}_{n=1}^{N'}$, that is assumed to have been independently drawn from the same distribution as the training observations. In this work, we consider NN classifiers that have already been trained on some training data and that achieve reasonable accuracies on test data following the same distribution as training data.

[1] While our exposition is restricted to fully-connected feedforward neural networks for the sake of concision, our representations are well-defined for other types of neural nets (*e.g.*, recurrent neural nets). In particular, they adapt seamlessly to the case of convolutional layers: such a layer can always be represented as a fully-connected layer whose weight matrix is constrained to have many zeroes, and what follows applies without further modifications.

3.3 Activation Graphs

Given an instance $x = x_0 \in \mathbb{R}^d$, a trained neural network $f_1, \ldots, f_L$ with $x_{\ell+1} = f_\ell(x_\ell) = \sigma_\ell(W_\ell \cdot x_\ell + b_\ell)$ and a layer $f_\ell \colon \mathbb{R}^{n_\ell} \longrightarrow \mathbb{R}^{n_{\ell+1}}$, we can define a weighted graph, called the *activation graph* $G_\ell(x)$ of x for the layer f_ℓ, as follows. We let $V \coloneqq V_\ell \sqcup V_{\ell+1}$ be the disjoint union of the two sets $V_\ell = \{1, \ldots, n_\ell\}$ and $V_{\ell+1} = \{1, \ldots, n_{\ell+1}\}$. The edges are defined as $E \coloneqq V_\ell \times V_{\ell+1}$. To each edge $(i, j) \in E_\ell$, we associate the weight $w_{i,j}(x) \coloneqq W_\ell(j, i) \cdot x_\ell(i)$, where $x_\ell(i)$ (resp. $W_\ell(j, i)$) denotes the i-th coordinate of $x_\ell \in \mathbb{R}^{n_\ell}$ (resp. entry (j, i) of $W_\ell \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$). The activation graph $G_\ell(x)$ is the weighted graph $(V, E, \{w_{i,j}(x)\})$, which can be conveniently represented as a $n_\ell \times n_{\ell+1}$ matrix whose entry (i, j) is $w_{i,j}(x)$. A simple illustration of this definition can be found in the Supplementary Material. Intuitively, these activation graphs—first considered in Gebhart et al. (2019)—represent how the network "reacts" to a given observation x at inner-level, rather than only considering the network output (*i.e.*, the Confidence Vector).

4 Two-Sample Statistical Tests using `MAGDiff`

4.1 The `MAGDiff` representations

Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two distributions for which we want to test $H_0 : \mathbb{P}_0 = \mathbb{P}_1$. As mentioned above, two-sample statistical tests tend to underperform when used directly on high-dimensional data. It is thus common practice to extract lower-dimensional representations $\Psi(x)$ from the data[2] $x \sim \mathbb{P}_i$, where $\Psi \colon \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1 \to \mathbb{R}^M$. Given a classification task with classes $1, \ldots, D$, we define a family of such representations as follows. Let $T \colon \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1 \to V$ be any map whose codomain V is a Banach space with norm $\| \cdot \|_V$. For each class $i \in \{1, \ldots, D\}$, let $\mathbb{P}_{0,i}$ be the conditional distribution of data points from $\mathbb{P}_0$ in class i. We define

$$\Psi_i(x) \coloneqq \|T(x) - \mathbb{E}_{\mathbb{P}_{0,i}}[T(x')]\|_V$$

for $x \in \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1$. Given a fixed finite dataset $x_1, \ldots, x_m \overset{\text{iid}}{\sim} \mathbb{P}_0$, we similarly define the approximation

$$\tilde{\Psi}_i(x) \coloneqq \|T(x) - \frac{1}{m_i} \sum_{j=1}^{m_i} T(x_j^i)\|_V,$$

where $x_1^i, \ldots, x_{m_i}^i$ are the points whose class is i. This defines a map $\tilde{\Psi} \colon \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1 \to \mathbb{R}^D$.

The map $T \colon \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1 \to V$ could *a priori* take many shapes. In this article, we assume that we are provided with a neural network F that has been trained for the classifying task at hand, as well as a training set drawn from $\mathbb{P}_0$. We let T be the activation graph G_ℓ of the layer f_ℓ of F represented as a matrix, so that the expected values $\mathbb{E}_{\mathbb{P}_{0,i}}[G_\ell(x')]$ (for $i = 1, \ldots, D$) are simply mean matrices, and the norm $\| \cdot \|_V$ is the Frobenius norm $\| \cdot \|_2$. We call the resulting D-dimensional representation *Mean Activation Graph Difference* (`MAGDiff`):

$$\texttt{MAGDiff}(x)_i \coloneqq \|G_\ell(x) - \frac{1}{m_i} \sum_{j=1}^{m_i} G_\ell(x_j^i)\|_2,$$

for $i = 1, \ldots, D$, where $x_1^i, \ldots, x_{m_i}^i$ are, as above, samples of the training set whose class is i. Therefore, for a given new observation x, we derive a vector $\texttt{MAGDiff}(x) \in \mathbb{R}^D$ whose i-th coordinate indicates whether x activates the chosen layer of the network in a similar way "as training observations of the class i".

Many variations are possible within that framework. One could, *e.g.*, consider the activation graph of several consecutive layers, use another matrix norm, or apply Topological Data Analysis techniques to compute a more compact representation of the graphs, such as the *topological uncertainty* (Lacombe et al., 2021). In this work, we focus on `MAGDiff` for dense layers, though it could be extended to other types.

[2]Here, as in the remainder of the article, we commit a minor abuse of notation: $\mathbb{P}_0$ and $\mathbb{P}_1$ are distributions on both the features and the labels, *i.e.* $(x, y) \sim \mathbb{P}_i$, but we often write $x \sim \mathbb{P}_i$ to indicate that x has been drawn from $(\mathbb{P}_i)_x$, the marginal of $\mathbb{P}_i$ with respect to the features. To avoid any confusion, we always let the letter x (possibly with a subscript) indicate features.

4.2 Comparison of distributions of features with multiple KS tests

Given as above a (relatively low-dimensional) representation $\Psi\colon \operatorname{supp} \mathbb{P}_0 \cup \operatorname{supp} \mathbb{P}_1 \to \mathbb{R}^N$ and samples $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathbb{P}_0$ and $x'_1, \ldots, x'_m \overset{\text{iid}}{\sim} \mathbb{P}_1$, one can apply multiple univariate (coordinate-wise) KS tests with Bonferroni correction to the sets $\Psi(x_1), \ldots, \Psi(x_n)$ and $\Psi(x'_1), \ldots, \Psi(x'_m)$, as described in Section 3. If Ψ is well-chosen, a difference between the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ (hard to test directly due to the dimensionality of the data) will translate to a difference between the distributions $\Psi(x)$ and $\Psi(x')$ for $x \sim \mathbb{P}_0$ and $x' \sim \mathbb{P}_1$ respectively. Detecting such a difference serves as a proxy for testing $H_0 : \mathbb{P}_0 = \mathbb{P}_1$. In our experiments, we apply this procedure to the `MAGDiff` representations defined above (see Section 5.1 for a step-by-step description). This is a reasonable approach, as it is a simple fact that a generic shift in the distribution of the random variable $x \sim \mathbb{P}_0$ will in turn induce a shift in the distribution of $\Psi(x)$, as long as Ψ is not constant[3]; however, this does not give us any true guarantee, as it does not provide any quantitative result regarding the shift in the distribution of $\Psi(x)$. Such results are beyond the scope of this paper, in which we focus on the good experimental performance of the `MAGDiff` statistic.

4.3 Differences from BBSD and motivations

The BBSD method described in Lipton et al. (2018) and Rabanser et al. (2019) is defined in a similar manner, except that the representations Ψ on which the multiple univariate KS tests are applied are simply the Confidence Vectors (CV) $F(x) \in \mathbb{R}^D$ of the neural network F (or of any other classifier that outputs confidence vectors), rather than our newly proposed `MAGDiff` representations. In other words, they detect shifts in the distribution of the inputs by testing for shifts in the distribution of the outputs of a given classifier[4].

Both our method and theirs share advantages: the features are task- and data-driven, as they are derived from a classifier that was trained for the specific task at hand. They do not require the design or the training of an additional model specifically geared towards shift detection, and they have favorable algorithmic complexity, especially compared to some kernel-based methods. In particular, combining the KS tests with the Bonferroni correction spares us from having to calibrate our statistical tests with a permutation test, which can be costly as shown in Rabanser et al. (2019). A common downside is that the Bonferroni correction can be overly conservative; other tests might offer higher power. The main focus of this article is the relevance of the `MAGDiff` representations, rather than the statistical tests that we apply to them, and it has been shown in Rabanser et al. (2019) that KS tests yield state-of-the-art performances; as such, we did not investigate alternatives, though additional efforts in that direction might produce even better results.

The nature of the construction of the `MAGDiff` representations is geared towards shift detection since it is directly based on encoding differences (*i.e.*, deviations) from the mean activation graphs (of $\mathbb{P}_0$). Moreover, they are based on representations from deeper within the NN, which are less compressed than the CV - passing through each layer leads to a potential loss of information. Hence, we can hope for the `MAGDiff` to encode more information from the input data than the CV representations used in Rabanser et al. (2019) which focus on the class to which a sample belongs to, while sharing the same favorable dimensionality reduction properties. Therefore, we expect `MAGDiff` to perform particularly well with covariate shifts, where shifts in the distribution of the data do not necessarily translate to strong shifts in the distribution of the CV. Conversely, we do not hope for our representations to bring significant improvements over CV in the specific case of label shifts; all the information relative to labels available to the network is, in a sense, best summarized in the CV, as this is the main task of the NN. These expectations were confirmed in our experiments.

5 Experiments

This experimental section is devoted to showcasing the use of the `MAGDiff` representations and its benefits over the well-established baseline CV when it comes to performing covariate shift detection. As detailed in

[3]See the Supplementary Material, Section 3 for an elementary proof.

[4]This corresponds to the best-performing variant of their method, denoted as *BBSDs* (as opposed to, *e.g.*, *BBSDh*) in Rabanser et al. (2019).

Section 5.1, we combine coordinate-wise KS tests for both these representations. Note that in the case of CV, this corresponds exactly to the method termed *BBSDs* in Rabanser et al. (2019). Our code is provided in the Supplementary Material, as well as a more thorough presentation of the datasets and parameters used.

5.1 Experimental Settings

Datasets. We consider the standard datasets MNIST (LeCun et al., 1998), FashionMNIST (FMNIST) (Xiao et al., 2017), CIFAR-10 (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), as well as a lighter version of ImageNet (restricted to 10 classes) called Imagenette (Ima, 2023).

Architectures. For MNIST and FMNIST, we used a simple CNN architecture consisting of 3 convolutional layers followed by 4 dense layers. For CIFAR-10 and SVHN, we considered (a slight modification, to account for input images of size 32×32, of) the ResNet18 architecture (He et al., 2015). For Imagenette, we used a pretrained ResNet18 model provided by Pytorch (Res, 2018). With these architectures, we reached a test accuracy of 98.6% on MNIST, 91.1% on FMNIST, 94.1% on SVHN, 81% on CIFAR-10 and 99.2% for Imagenette, validating the "well-trained" assumption mentioned in Section 4. Note that we used simple architectures, without requiring the networks to achieve state-of-the-art accuracy.

Shifts. We applied three types of shift to our datasets: Gaussian noise (additive white noise), Gaussian blur (convolution by a Gaussian distribution), and Image shift (random combination of rotation, translation, zoom and shear), for six different levels of increasing intensities (denoted by I, II,...,VI), and a fraction of shifted data $\delta \in \{0.25, 0.5, 1.0\}$. For each dataset and shift type, we chose the shift intensities in such a manner that the shift detection for the lowest intensities and low δ is almost indetectable for both methods (`MAGDiff` and CV), and very easily detectable for high intensities and values of δ. Details (including the impact of the shifts on model accuracy) and illustrations can be found in the Supplementary Material.

Sample size. We ran the shift detection tests with sample sizes[5] $\{10, 20, 50, 100, 200, 500, 1000\}$ to assess how many samples a given method requires to reliably detect a distribution shift. A good method should be able to detect a shift with as few samples as possible.

Experimental protocol. In all of the experiments below, we start with a neural network that is pre-trained on the training set of a given dataset. The test set will be referred to as the *clean set* (CS). We then apply the selected shift (type, intensity, and proportion δ) to the clean set and call the resulting set the *shifted set* SS; it represents the target distribution $\mathbb{P}_1$ in the case where $\mathbb{P}_1 \neq \mathbb{P}_0$.

As explained in Section 4, for each of the classes $i = 1, \ldots, D$ (for all of our datasets, $D = 10$), we compute the mean activation graph of a chosen dense layer f_ℓ of (a random subset of size 1000 of all) samples in the training set whose class is i; this yields D mean activation graphs $G_1, \ldots, G_D$. We compute for each sample x in CS and each sample in SS the representation $\texttt{MAGDiff}(x)$, where $\texttt{MAGDiff}(x)_i = \|G_\ell(x) - G_i\|_2$ for $i = 1, \ldots, D$ and $G_\ell(x)$ is the activation graph of x for the layer f_ℓ (as explained in Section 4). Doing so, we obtain two sets $\{\texttt{MAGDiff}(x) \mid x \in CS\}$ and $\{\texttt{MAGDiff}(x') \mid x' \in SS\}$ of D-dimensional features with the same cardinality as the test set.

Now, we estimate the power of the test for a given sample size[6] m and for a type I error of at most 0.05; in other words, the probability that the test rejects H_0 when H_1 is true and when it has access to only m samples from the respective datasets, and under the constraint that it does not falsely reject H_0 in more than 5% of cases. To do so, we randomly sample (with replacement) m elements $x'_1, \ldots, x'_m$ from SS, and consider for each class $i = 1, \ldots, D$ the discrete empirical univariate distribution q_i of the values $\texttt{MAGDiff}(x'_1)_i, \ldots, \texttt{MAGDiff}(x'_m)_i$. Similarly, by randomly sampling m elements from CS, we obtain another discrete univariate distribution p_i (see Figure 1 for an illustration). Then, for each $i = 1, \ldots, D$, the KS test is used to compare p_i and q_i to obtain a p-value λ_i, and reject H_0 if $\min(\lambda_1, \ldots, \lambda_D) < \alpha/D$, where α is the threshold for the univariate KS test at confidence 0.05 (*cf.* Section 3.1). Following standard bootstrapping

[5] That is, the number of elements from the clean and shifted sets on which the statistical tests are performed; see the paragraph **Experimental protocol** for more details.

[6] The same sample size that is mentioned in the **Sample size** paragraph.

protocol, we repeat that experiment (independently sampling m points from CS and SS, computing p-values, and possibly rejecting H_0) 1500 times; the percentage of rejection of H_0 is the estimated *power* of the statistical test (since H_0 is false in this scenario). We use the asymptotic normal distribution of the standard Central Limit Theorem to compute approximate 95%-confidence intervals on our estimate.

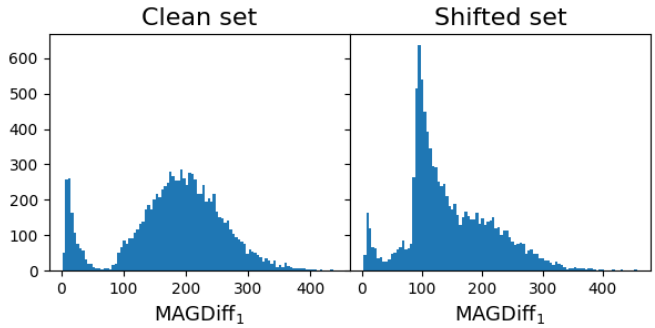

Figure 1: Empirical distributions of $\texttt{MAGDiff}_1$ for the $10,000$ samples of the clean and shifted sets (MNIST, Gaussian noise, $\delta = 0.5$, last dense layer). For the clean set, the distribution of the component $\texttt{MAGDiff}_1$ of `MAGDiff` exhibits a peak close to 0. This corresponds to those samples whose distance to the mean activation graph of (training) samples belonging to the associated class is very small, *i.e.*, these are samples that presumably belong to the same class as well. Note that, for the shifted set, this peak close to 0 is substantially diminished, which indicates that the activation graph of samples affected by the shift is no longer as close to the mean activation graph of their true class.

To illustrate that the test is well calibrated, we repeat the same procedure while sampling twice m elements from CS (rather than m elements from SS and m elements from CS), which allows us to estimate the type I error (*i.e.*, the percentage of incorrect rejections of H_0) and assert that it remains below the significance level of 5% (see, *e.g.*, Figure 2).

We experimented with a few variants of the `MAGDiff` representations: we tried reordering the coordinates of each vector $\texttt{MAGDiff}(x) \in \mathbb{R}^D$ in increasing order of the value of the associated confidence vectors. We also tried replacing the matrix norm of the difference to the mean activation graph by either its Topological Uncertainty (TU) (Lacombe et al., 2021), or variants thereof. Early analysis suggested that these variations did not bring increased performances, despite their increased complexity. Experiments also suggested that `MAGDiff` representations brought no improvement over CV in the case of label shift. We also tried to combine (*i.e.*, concatenate) the CV and `MAGDiff` representations, but the results were unimpressive, which we attribute to the Bonferroni correction getting more conservative the higher the dimension. We thus only report the results for the standard `MAGDiff`.

Competitor. We used multiple univariate KS tests applied to CV (the method BBSDs from Rabanser et al. (2019)) as the baseline, which we denote by "CV" in the figures and tables, in contrast to our method denoted by "`MAGDiff`". The similarity in the statistical testing between BBSDs and `MAGDiff` allows us to easily assess the relevance of the `MAGDiff` features. We chose them as our sole competitors as it has been convincingly shown in Rabanser et al. (2019) that they outperform on average all other standard methods, including the use of dedicated dimensionality reduction models, such as autoencoders, or of multivariate kernel tests. Many of these methods are also either computationally more costly (to the point where they cannot be practically applied to more than a thousand samples) or harder to implement (as they require an additional neural network to be implemented) than both BBSDs and `MAGDiff`.

5.2 Experimental results and influence of parameters.

We now showcase the power of shift detection using our `MAGDiff` representations in various settings and compare it to the state-of-the-art competitor CV. Since there were a large number of hyper-parameters in our experiments (datasets, shift types, shift intensities, etc.), we started with a standard set of hyper-parameters that yielded representative and informative results according to our observations (MNIST and Gaussian noise, as in Rabanser et al. (2019), $\delta = 0.5$, sample size 100, `MAGDiff` computed with the last layer of the network) and let some of them vary in the successive experiments. We focus on the well-known MNIST

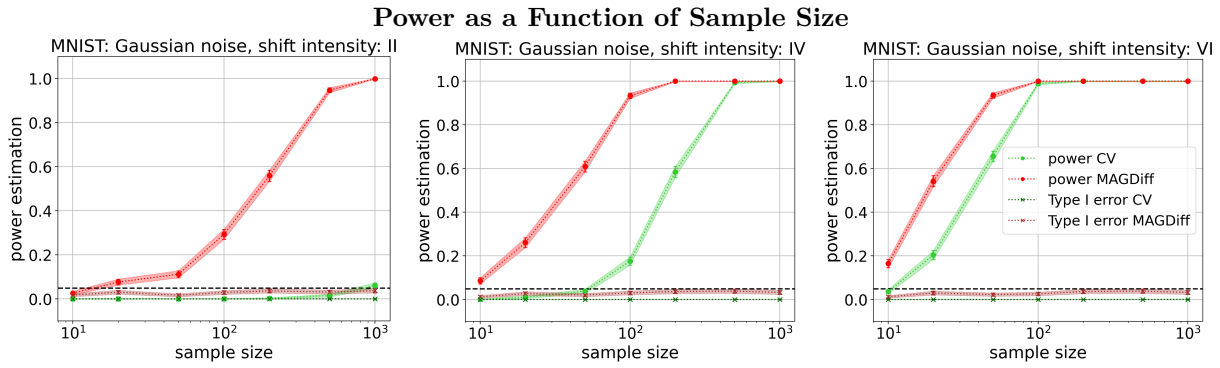

Figure 2: Power and type I error of the statistical test with `MAGDiff` (red) and CV (green) representations w.r.t. sample size (on a log-scale) for three different shift intensities (II, IV, VI) and fixed $\delta = 0.5$ for the MNIST dataset, Gaussian noise and last layer of the network, with estimated 95%-confidence intervals.

dataset to allow for easy comparisons, and refer to the Supplementary Material for additional experimental results that confirm our findings on other datasets.

Sample size. The first experiment consists of estimating the power of the shift detection test as a function of the sample size (a common way of measuring the performance of such a test) using either the `MAGDiff` or the baseline CV representations. Figure 2 shows the powers of the KS tests using the `MAGDiff` (red curve) and CV (green curve) representations with respect to the sample size for the MNIST dataset. Here, we choose to showcase the results for Gaussian noise of intensities II, IV and IV with shift proportion $\delta = 0.5$.

It can clearly be seen that `MAGDiff` consistently and significantly outperformed the CV representations. While in both cases, the tests achieved a power of 1.0 for large sample sizes ($m \approx 1000$) and/or high shift intensity (VI), `MAGDiff` was capable of detecting the shift even with much lower sample sizes. This was particularly striking for the low intensity level II, where the test with CV was completely unable to detect the shift, even with the largest sample size, while `MAGDiff` was capable of reaching non-trivial power already for a medium sample size of 100 and exceptional power for large sample size. Note that the tests were always well-calibrated. That is, the type I error remained below the significance level of 0.05, indicated by the horizontal dashed black line in the figures.

To further support our claim that `MAGDiff` outperforms CV on average in other scenarios, we provide, in Table 1, averaged results over all parameters except the sample size. Though the precise values obtained are not particularly informative (due to the aggregation over very different sets of hyper-parameters), the comparison between the two rows remains relevant. In the Supplementary Material, a more comprehensive experimental report (including, in particular, the CIFAR-10 and Imagenette datasets) further supports our claims.

	Averaged power (%)						
Sample size	10	20	50	100	200	500	1000
`MAGDiff`	7.4	17.1	27.6	40.7	54.7	71.4	80.4
CV	4.0	9.8	15.6	24.7	35.3	49.7	59.2

Table 1: Averaged test power of `MAGDiff` and CV over all hyper-parameters except sample size (dataset, shift type, δ, shift intensity). A 95%-confidence interval for the averaged powers has been estimated via bootstrapping and is, in all cases, strictly contained in a $\pm 0.1\%$ interval.

Shift intensity. The first experiment suggests that `MAGDiff` representations perform particularly well when the shift is hard to detect. In the second experiment, we further investigate the influence of the shift intensity level and δ (which is, in a sense, another measure of shift intensity) on the power of the tests. We chose a fixed sample size of 100, which was shown to make for a challenging yet doable task. The results in Figure 3 confirm that our representations were much more sensitive to weak shifts than the CV, with differences in power greater than 80% for some intensities.

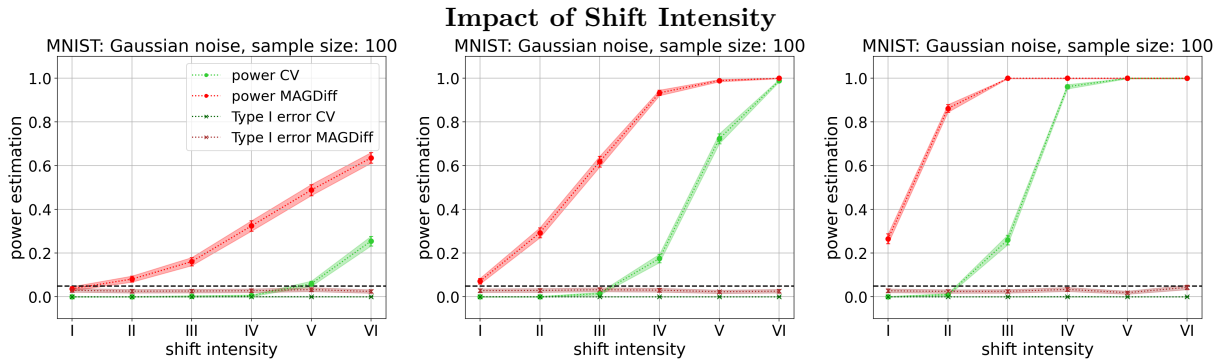

Figure 3: Power and type I error of the test with `MAGDiff` (red) and CV (green) features w.r.t. the shift intensity for Gaussian noise on the MNIST dataset with sample size 100 and $\delta = 0.25$ (left), $\delta = 0.5$ (middle), $\delta = 1.0$ (right), for the last dense layer. The estimated 95%-confidence intervals are displayed around the curves.

Impact of Shift Type

Shift	Feat.	Power of the test (%) — Shift intensity I	II	III	IV	V	VI
GN	`MD`	7.2 ± 1.3	29.3 ± 2.3	61.9 ± 2.5	93.3 ± 1.3	98.9 ± 0.5	$100.0 - 0.0$
	CV	$0.0 + 0.2$	0.1 ± 0.1	1.5 ± 0.6	17.6 ± 1.9	72.3 ± 2.3	98.9 ± 0.5
GB	`MD`	3.7 ± 1.0	4.3 ± 1.0	27.7 ± 2.3	63.1 ± 2.4	85.0 ± 1.8	92.4 ± 1.3
	CV	$0.0 + 0.2$	$0.0 + 0.2$	$0.0 + 0.2$	0.4 ± 0.3	1.3 ± 0.6	5.3 ± 1.1
IS	`MD`	10.3 ± 1.5	32.7 ± 2.4	53.5 ± 2.5	78.5 ± 2.1	90.6 ± 1.5	98.9 ± 0.5
	CV	$0.0 + 0.2$	0.1 ± 0.2	2.1 ± 0.7	11.5 ± 1.6	37.0 ± 2.4	86.3 ± 1.7

Table 2: Power of the two methods (our method, denoted as `MD`, and CV) as a function of the shift intensity for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS) on the MNIST dataset with $\delta = 0.5$, sample size 100, for the last dense layer. Red indicates that the estimated power is below 10%, blue that it is above 50%. The 95%-confidence intervals have been estimated as mentioned in Section 5.

Shift type. The previous experiments focused on the case of Gaussian noise; in this experiment, we investigate whether the results hold for other shift types. As detailed in Table 2, the test with `MAGDiff` representations reacted to the shifts even for low intensities of I, II, and III for all shift types (Gaussian blur being the most difficult case), while the KS test with CV was unable to detect anything. For medium to high intensities III, IV, V and VI, `MAGDiff` again significantly outperformed the baseline and reaches powers close to 1 for all shift types. For the Gaussian blur, the shift remained practically undetectable using CV.

`MAGDiff` with respect to different layers. The NN architecture we used with MNIST and FMNIST had several dense layers before the output. As a variation of our method, we investigate the effect on the shift detection when computing our `MAGDiff` representations with respect to different layers[7]. More precisely, we consider the last three dense layers denoted by ℓ_{-1}, ℓ_{-2} and ℓ_{-3}, ordered from the closest to the network output (ℓ_{-1}) to the third from the end (ℓ_{-3}). The averaged results over all parameters and noise types are in Table 3.

In the case of MNIST we only observe a slight increase in power when considering layer ℓ_{-3} further from the output of

Choice of Layer

Dataset	Averaged power (%) — Features CV	ℓ_{-1}	ℓ_{-2}	ℓ_{-3}
MNIST	25.1	51.9	53.0	56.4
FMNIST	46.2	44.9	47.6	53.7

Table 3: Averaged performance of the various layers for `MAGDiff` over all other parameters (for MNIST and FMNIST), compared to BBSD with CV. A 95% confidence interval for the averaged powers was estimated and is in all cases contained in a $\pm 0.1\%$ interval.

[7]Since ResNet18 only has a single dense layer after its convolutional layers, there is no choice to be made in the case of CIFAR-10, SVHN and Imagenette.

the NN. In the case of FMNIST, on the other hand, we clearly see a much more pronounced improvement when switching from ℓ_{-1} to ℓ_{-3}. This hints at the possibility that features derived from encodings further within the NN can, in some cases, be more pertinent to the task of shift detection than those closer to the output.

6 Conclusion

In this article, we derive new representations `MAGDiff` from the activation graphs of a trained NN classifier. We empirically show that using `MAGDiff` representations for data set shift detection via coordinate-wise KS tests (with Bonferroni correction) significantly outperforms the baseline given by using confidence vectors established in Lipton et al. (2018), while remaining equally fast and easy to implement, making `MAGDiff` representations an efficient tool for this critical task.

Our findings open many avenues for future investigations. We focused on classification of image data in this work, but our method is a general one and can be applied to other settings. Moreover, adapting our method to regression tasks as well as to settings where shifts occur gradually is feasible and a starting point for future work. Finally, exploring variants of the `MAGDiff` representations—considering several layers of the network at once, extending it to other types of layers, extracting finer topological information from the activation graphs, weighting the edges of the graph by backpropagating their contribution to the output, etc.—could also result in increased performance.

References

Pre-trained weights for resnet18. `https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`, 2018. Accessed: 10/05/2023.

Imagenette dataset. `https://github.com/fastai/imagenette`, 2023. Accessed: 10/05/2023.

Florence Alberge, Clément Feutry, Pierre Duhamel, and Pablo Piantanida. Detecting covariate shift with Black Box predictors. In *International Conference on Telecommunications (ICT 2019)*, Hanoi, Vietnam, April 2019. doi: 10.1109/ICT.2019.8798827. URL `https://hal-centralesupelec.archives-ouvertes.fr/hal-02172275`.

Guy Bar-Shalom, Yonatan Geifman, and Ran El-Yaniv. Distribution shift detection for deep neural networks, 2022. URL `https://arxiv.org/abs/2210.10897`.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41, 07 2009. doi: 10.1145/1541880.1541882.

Thomas Gebhart, Paul Schrater, and Alan Hylton. Characterizing the shape of activation space in deep neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1537–1542. IEEE, 2019.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

James Heckman. Sample selection bias as a specification error (with an application to the estimation of labor supply functions). *National Bureau of Economic Research, Inc, NBER Working Papers*, 01 1977.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

C. Heumann and M. Schomaker. *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*. Springer International Publishing, 2023. ISBN 9783031118333. URL https://books.google.fr/books?id=DrSqEAAAQBAJ.

Vitor A.C. Horta, Ilaria Tiddi, Suzanne Little, and Alessandra Mileo. Extracting knowledge from deep neural networks through graph analysis. *Future Generation Computer Systems*, 120:109–118, 2021. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2021.02.009. URL https://www.sciencedirect.com/science/article/pii/S0167739X21000613.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani. Sequential covariate shift detection using classifier two-sample tests. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9845–9880. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jang22a.html.

Gopinath Kallianpur. The topology of weak convergence of probability measures. *Journal of Mathematics and Mechanics*, pp. 947–969, 1961.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Théo Lacombe, Yuichi Ike, Mathieu Carrière, Frédéric Chazal, Marc Glisse, and Yuhei Umeda. Topological Uncertainty: monitoring trained neural networks through persistence of activation graphs. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pp. 2666–2672. International Joint Conferences on Artificial Intelligence Organization, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *35th International Conference on Machine Learning (ICML 2018)*, volume 80, pp. 3122–3130. PMLR, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2019)*, pp. 1396–1408. Curran Associates, Inc., 2019.

Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry A. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI Conference on Artificial Intelligence*, 2014.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2, 06 2012.

Alireza Shafaei, Mark W. Schmidt, and J. Little. Does your model know the digit 6 is not a cat? a less biased evaluation of "outlier" detectors. *ArXiv*, abs/1809.04729, 2018.

Nikolai V Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14, 1939.

Amos Storkey. When training and test sets are different: Characterizing learning transfer. *Dataset Shift in Machine Learning*, pp. 3–28, 01 2009. doi: 10.7551/mitpress/9780262170055.003.0001.

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems*, 33:49–61, 2020.

Simon Voss and Steve George. Multiple significance tests. *BMJ*, 310(6986):1073, 1995. ISSN 0959-8138. doi: 10.1136/bmj.310.6986.1073. URL `https://www.bmj.com/content/310/6986/1073.1`.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift, 2021a. URL `https://arxiv.org/abs/2110.11328`.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL `http://arxiv.org/abs/1708.07747`. cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/a49e9411d64ff53eccfdd09ad10a15b3-Paper.pdf`.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–819–III–827. JMLR.org, 2013.

A Appendix

A.1 Additional Information on the Experimental Procedures.

Datasets. The number of samples in the clean sets (*i.e.*, the test sets) of the datasets we investigated are as follows:

- MNIST, FMNIST and CIFAR-10: $10'000$,
- SVHN: $26'032$.
- Imagenette: $3'925$.

A sample from each dataset can be seen in Figure 4.

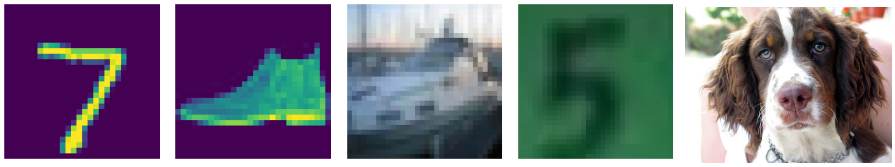

Figure 4: Sample images from all datasets used in the paper. From left to right: MNIST, FashionMNIST, CIFAR-10, SVHN and Imagenette.

Shifts. In order to illustrate the effect of the shift types described in Section 5.1 of the main article, we show the effects of the shifts and their intensities on the MNIST dataset in Figure 5. For the detailed parameters of each shift intensity (per dataset) we refer to the associated code.

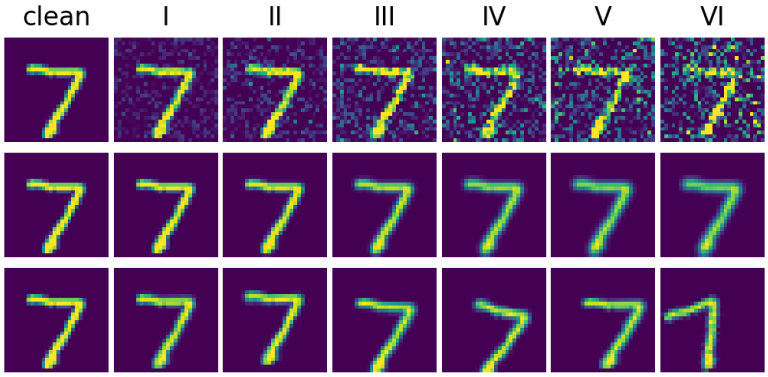

Figure 5: Illustration of intensities of the shift types — Gaussian noise (top row), Gaussian blur (middle row) and Image shift (bottom row) — on a sample from the MNIST dataset.

A remark on the experimental setup As explained in Subsection 5.1 of the main text, we perform each of our statistical tests on random subsamples from a clean set CS and a shifted set SS that has been crafted from CS by applying a controlled shift to its elements. We could alternatively have crafted SS by applying the shift to another "clean" set drawn from the same distribution as CS. The advantage of our current setup is that it ensures that the only difference between the distributions of CS and SS comes from the shift, which we completely control, as opposed to some (small) pre-shift difference between the two clean sets that could arise from the sampling process in this alternative setup. In practice, however, both setups should yield extremely similar results; indeed, we randomly subsample from CS and SS to get the sets on which we perform the tests, and those sets are typically of size ~ 100, whereas CS and SS are of size $3'925$, $10'000$ or $26'032$ depending on the dataset. Hence the subsampled sets from the CS and from SS should have almost no element in common (by which we mean the same picture appearing shifted in one of the sets and unshifted in the other), as in the other setup.

An example of activation graph computation We illustrate the definition of the activation graph from Subsection 3.3, using the same notations. A layer $f_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_{\ell+1}}$ is shown on the left of Figure 6. Here $n_\ell = 2$ and $n_{\ell+1} = 3$, and the weight matrix associated to f_ℓ is $W_\ell = \begin{bmatrix} -1 & 0 \\ 2 & 4 \\ 3 & -2 \end{bmatrix}$. Given an input x to the network that corresponds to an input $x_\ell = (1, -2)$ to the layer f_ℓ, the associated activation graph $G_\ell(x)$ is shown on the right of Figure 6.

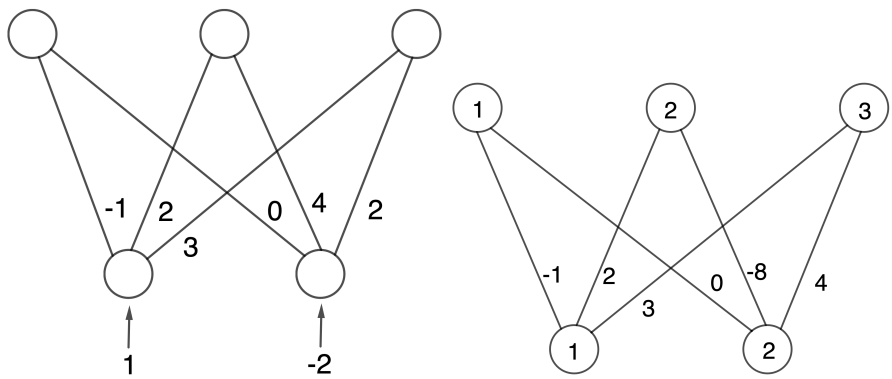

Figure 6: On the left, a neural network layer f_ℓ with the associated weights and an input vector x_ℓ. On the right, the associated activation graph.

Shift	Int.	Feat.	MNIST						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	2.7 ± 0.8	7.7 ± 1.3	11.2 ± 1.6	29.3 ± 2.3	55.9 ± 2.5	94.7 ± 1.1	99.9 ± 0.1
		CV	$0.0 + 0.2$	0.1 ± 0.1	$0.0 + 0.2$	0.1 ± 0.1	0.3 ± 0.3	1.5 ± 0.6	6.1 ± 1.2
	IV	MD	8.6 ± 1.4	26.1 ± 2.2	60.9 ± 2.5	93.3 ± 1.3	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$
		CV	0.1 ± 0.2	1.0 ± 0.5	3.8 ± 1.0	17.6 ± 1.9	58.4 ± 2.5	99.3 ± 0.4	$100.0 - 0.0$
	VI	MD	16.5 ± 1.9	54.2 ± 2.5	93.5 ± 1.3	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$
		CV	3.7 ± 1.0	20.4 ± 2.0	65.6 ± 2.4	98.9 ± 0.5	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$
GB	II	MD	1.9 ± 0.7	2.7 ± 0.8	3.3 ± 0.9	4.3 ± 1.0	9.9 ± 1.5	22.2 ± 2.1	40.7 ± 2.5
		CV	$0.0 + 0.2$	0.1 ± 0.1	$0.0 + 0.2$	$0.0 + 0.2$	$0.0 + 0.2$	0.1 ± 0.1	0.1 ± 0.1
	IV	MD	4.8 ± 1.1	13.1 ± 1.7	30.3 ± 2.3	63.1 ± 2.4	93.9 ± 1.2	$100.0 - 0.0$	$100.0 - 0.0$
		CV	$0.0 + 0.2$	$0.0 + 0.2$	0.1 ± 0.2	0.3 ± 0.3	1.5 ± 0.6	11.3 ± 1.6	44.9 ± 2.5
	VI	MD	9.2 ± 1.5	25.1 ± 2.2	57.5 ± 2.5	92.4 ± 1.3	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$
		CV	0.1 ± 0.1	0.5 ± 0.4	1.4 ± 0.6	5.1 ± 1.1	22.1 ± 2.1	88.5 ± 1.6	$100.0 - 0.0$
IS	II	MD	3.5 ± 0.9	8.6 ± 1.4	15.1 ± 1.8	32.7 ± 2.4	66.3 ± 2.4	98.0 ± 0.7	$100.0 - 0.0$
		CV	$0.0 + 0.2$	$0.0 + 0.2$	$0.0 + 0.2$	0.1 ± 0.2	0.7 ± 0.4	6.9 ± 1.3	28.4 ± 2.3
	IV	MD	5.6 ± 1.2	18.5 ± 2.0	42.1 ± 2.5	78.5 ± 2.1	98.0 ± 0.7	$100.0 - 0.0$	$100.0 - 0.0$
		CV	0.1 ± 0.2	0.9 ± 0.5	2.4 ± 0.8	15.5 ± 1.8	50.0 ± 2.5	99.5 ± 0.3	$100.0 - 0.0$
	VI	MD	10.4 ± 1.5	34.0 ± 2.4	72.6 ± 2.3	98.9 ± 0.5	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$
		CV	1.3 ± 0.6	7.3 ± 1.3	31.7 ± 2.4	83.3 ± 1.9	$100.0 - 0.0$	$100.0 - 0.0$	$100.0 - 0.0$

Table 4: Power of the statistical test with `MAGDiff` (abbreviated as `MD`) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed $\delta = 0.5$ for the MNIST dataset. The estimated 95%-confidence intervals are indicated.

A.2 Additional Experimental Results

Sample size. To further support our claims, we include comprehensive results of the power with respect to the sample size for the MNIST, Imagenette and CIFAR-10 datasets in Tables 4, 5 and 6. We provide all results for the shift intensities II, IV and VI, for all shift types, and fixed $\delta = 0.5$ for MNIST, CIFAR-10, respectively $\delta = 1.0$ for Imagenette (the δ were chosen so that the task is comparatively easy at high shift intensity and hard at low shift intensity for both methods).

Shift intensity. In Figures 7, 8 and 9, we collect the plots of the estimated powers of the test for multiple cases, in addition to the one presented in the main article. Note that the only situation in which `MAGDiff` is very slightly outperformed by the baseline CV, is the case of FMNIST, when we consider `MAGDiff` representations of layer l_{-1}. In all other cases, shift detection using `MAGDiff` representations clearly outperforms the baseline of CV by a large margin.

Model accuracy. In Figure 10 we show the impact of the shift type and intensity on the model accuracy. It is interesting to note that, even in cases where the model accuracy is only minimally impacted (*e.g.*, for Gaussian blur on the MNIST and FMNIST datasets), our method can still reliably detect the presence of the shift.

Norm variations. As mentioned in the main paper, many variations of `MAGDiff` are conceivable. Here, we present some experimental results for variations on the type of norm that is used to construct the `MAGDiff` representations. In Figure 11 we show the results where, instead of the Frobenius-norm, we consider the *spectral norm* as well as the operator norm $\|\cdot\|_\infty$ induced by the sup-norm on vectors. The spectral norm is equal to the largest singular value and $\|\cdot\|_\infty$ is defined by:

$$\|M\|_\infty \coloneqq \sup_{x \neq 0} \frac{\|Mx\|_\infty}{\|x\|_\infty} = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |m_{ij}|$$

Shift	Int.	Feat.	Imagenette						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	0.7 ± 0.4	2.2 ± 0.7	6.3 ± 1.2	16.7 ± 1.9	46.2 ± 2.5	93.3 ± 1.3	99.9 ± 0.1
		CV	2.4 ± 0.8	4.5 ± 1.1	2.7 ± 0.8	4.1 ± 1.0	4.2 ± 1.0	7.3 ± 1.3	10.3 ± 1.5
	IV	MD	0.9 ± 0.5	3.6 ± 0.9	7.3 ± 1.3	22.1 ± 2.1	60.7 ± 2.5	98.5 ± 0.6	$100.0 - 0.0$
		CV	1.6 ± 0.6	4.1 ± 1.0	3.2 ± 0.9	3.9 ± 1.0	4.7 ± 1.1	7.5 ± 1.3	10.1 ± 1.5
	VI	MD	0.9 ± 0.5	3.6 ± 0.9	7.3 ± 1.3	22.1 ± 2.1	60.7 ± 2.5	98.5 ± 0.6	$100.0 - 0.0$
		CV	2.2 ± 0.7	4.9 ± 1.1	3.5 ± 0.9	5.3 ± 1.1	6.8 ± 1.3	15.5 ± 1.8	33.5 ± 2.4
GB	II	MD	0.5 ± 0.3	2.5 ± 0.8	4.1 ± 1.0	15.3 ± 1.8	40.6 ± 2.5	91.7 ± 1.4	99.9 ± 0.2
		CV	2.1 ± 0.7	3.7 ± 1.0	3.2 ± 0.9	3.7 ± 1.0	7.0 ± 1.3	11.3 ± 1.6	18.2 ± 2.0
	IV	MD	1.0 ± 0.5	3.5 ± 0.9	9.1 ± 1.5	29.3 ± 2.3	67.9 ± 2.4	99.1 ± 0.5	$100.0 - 0.0$
		CV	2.3 ± 0.8	4.5 ± 1.0	3.7 ± 1.0	4.5 ± 1.1	6.3 ± 1.2	13.1 ± 1.7	26.9 ± 2.2
	VI	MD	1.3 ± 0.6	5.0 ± 1.1	17.2 ± 1.9	50.5 ± 2.5	89.9 ± 1.5	$100.0 - 0.0$	$100.0 - 0.0$
		CV	2.1 ± 0.7	3.5 ± 0.9	3.1 ± 0.9	5.0 ± 1.1	6.9 ± 1.3	18.5 ± 2.0	46.8 ± 2.5
IS	II	MD	0.5 ± 0.4	1.1 ± 0.5	1.9 ± 0.7	4.7 ± 1.1	13.6 ± 1.7	44.5 ± 2.5	83.1 ± 1.9
		CV	2.5 ± 0.8	3.7 ± 1.0	3.9 ± 1.0	3.1 ± 0.9	4.3 ± 1.0	6.3 ± 1.2	9.9 ± 1.5
	IV	MD	0.3 ± 0.3	1.9 ± 0.7	2.9 ± 0.9	9.1 ± 1.5	28.1 ± 2.3	75.1 ± 2.2	98.3 ± 0.7
		CV	1.7 ± 0.7	3.0 ± 0.9	3.6 ± 0.9	3.9 ± 1.0	4.8 ± 1.1	8.3 ± 1.4	12.8 ± 1.7
	VI	MD	0.6 ± 0.4	2.5 ± 0.8	5.0 ± 1.1	14.9 ± 1.8	44.3 ± 2.5	93.5 ± 1.2	99.9 ± 0.1
		CV	2.0 ± 0.7	3.9 ± 1.0	1.9 ± 0.7	4.6 ± 1.1	6.1 ± 1.2	8.3 ± 1.4	15.5 ± 1.8

Table 5: Power of the statistical test with `MAGDiff` (abbreviated as `MD`) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed $\delta = 1$ for the Imagenette dataset. The estimated 95%-confidence intervals are indicated.

Shift	Int.	Feat.	CIFAR-10						
			Sample size						
			10	20	50	100	200	500	1000
GN	II	MD	1.8 ± 0.7	5.3 ± 1.1	16.7 ± 1.9	47.0 ± 2.5	86.9 ± 1.7	$100.0 - 0.0$	$100.0 - 0.0$
		CV	2.3 ± 0.8	4.5 ± 1.1	6.9 ± 1.3	19.1 ± 2.0	38.3 ± 2.5	88.3 ± 1.6	99.9 ± 0.1
	IV	MD	2.5 ± 0.8	11.1 ± 1.6	36.7 ± 2.4	81.1 ± 2.0	99.3 ± 0.4	$100.0 - 0.0$	$100.0 - 0.0$
		CV	2.5 ± 0.8	6.7 ± 1.3	11.7 ± 1.6	29.9 ± 2.3	63.4 ± 2.4	99.3 ± 0.4	$100.0 - 0.0$
	VI	MD	2.7 ± 0.8	14.7 ± 1.8	49.2 ± 2.5	91.2 ± 1.4	99.9 ± 0.1	$100.0 - 0.0$	$100.0 - 0.0$
		CV	2.7 ± 0.8	7.3 ± 1.3	14.2 ± 1.8	37.5 ± 2.5	77.3 ± 2.1	99.9 ± 0.2	$100.0 - 0.0$
GB	II	MD	0.8 ± 0.5	2.8 ± 0.8	6.6 ± 1.3	18.9 ± 2.0	49.5 ± 2.5	93.2 ± 1.3	$100.0 - 0.0$
		CV	2.6 ± 0.8	4.0 ± 1.0	3.4 ± 0.9	6.8 ± 1.3	11.1 ± 1.6	30.4 ± 2.3	58.9 ± 2.5
	IV	MD	1.8 ± 0.7	5.7 ± 1.2	19.5 ± 2.0	49.7 ± 2.5	89.6 ± 1.5	99.9 ± 0.1	$100.0 - 0.0$
		CV	2.5 ± 0.8	6.4 ± 1.2	7.3 ± 1.3	13.9 ± 1.7	35.9 ± 2.4	84.0 ± 1.9	99.8 ± 0.2
	VI	MD	2.1 ± 0.7	6.3 ± 1.2	23.4 ± 2.1	62.5 ± 2.4	96.1 ± 1.0	$100.0 - 0.0$	$100.0 - 0.0$
		CV	3.0 ± 0.9	8.9 ± 1.4	14.7 ± 1.8	44.2 ± 2.5	85.5 ± 1.8	$100.0 - 0.0$	$100.0 - 0.0$
IS	II	MD	0.3 ± 0.3	1.3 ± 0.6	3.6 ± 0.9	6.7 ± 1.3	19.6 ± 2.0	60.1 ± 2.5	92.6 ± 1.3
		CV	2.5 ± 0.8	3.3 ± 0.9	2.5 ± 0.8	4.4 ± 1.0	7.9 ± 1.4	16.5 ± 1.9	31.5 ± 2.4
	IV	MD	0.6 ± 0.4	2.4 ± 0.8	3.9 ± 1.0	16.1 ± 1.9	39.9 ± 2.5	88.2 ± 1.6	99.9 ± 0.1
		CV	2.0 ± 0.7	3.9 ± 1.0	2.9 ± 0.9	6.7 ± 1.3	10.8 ± 1.6	25.4 ± 2.2	53.1 ± 2.5
	VI	MD	1.3 ± 0.6	3.9 ± 1.0	8.9 ± 1.4	22.8 ± 2.1	57.4 ± 2.5	97.7 ± 0.8	$100.0 - 0.0$
		CV	2.0 ± 0.7	4.6 ± 1.1	4.4 ± 1.0	9.5 ± 1.5	15.5 ± 1.8	44.5 ± 2.5	83.2 ± 1.9

Table 6: Power of the statistical test with `MAGDiff` (abbreviated as `MD`) and CV representations for the shift types Gaussian noise (GN), Gaussian blur (GB) and Image shift (IS), three different shift intensities (II, IV, VI) and fixed $\delta = 0.5$ for the CIFAR-10 dataset. The estimated 95%-confidence intervals are indicated.

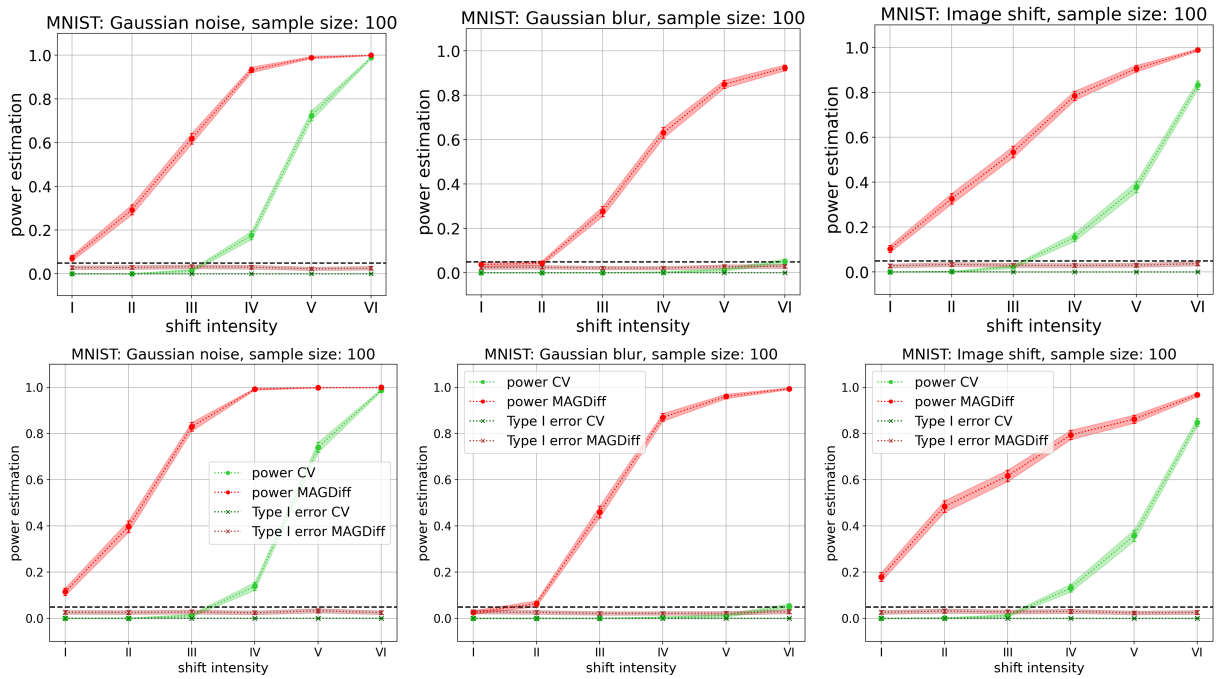

Figure 7: Power and type I error of the test with `MAGDiff` (red) and CV (green) representations w.r.t. the shift intensity for various shift types on the MNIST dataset with $\delta = 0.5$, sample size 100, for layers ℓ_{-1} (top row) and ℓ_{-3} (bottom row).

for $M \in \mathbb{R}^{m \times n}$. Comparing to Figure 7, we observe that the results for the Frobenius-norm and the spectral norm are almost identical. However, while the results for the $\|\cdot\|_\infty$ are still better (in almost all cases) than those of the baseline CV, they are less powerful than those of the Frobenius norm.

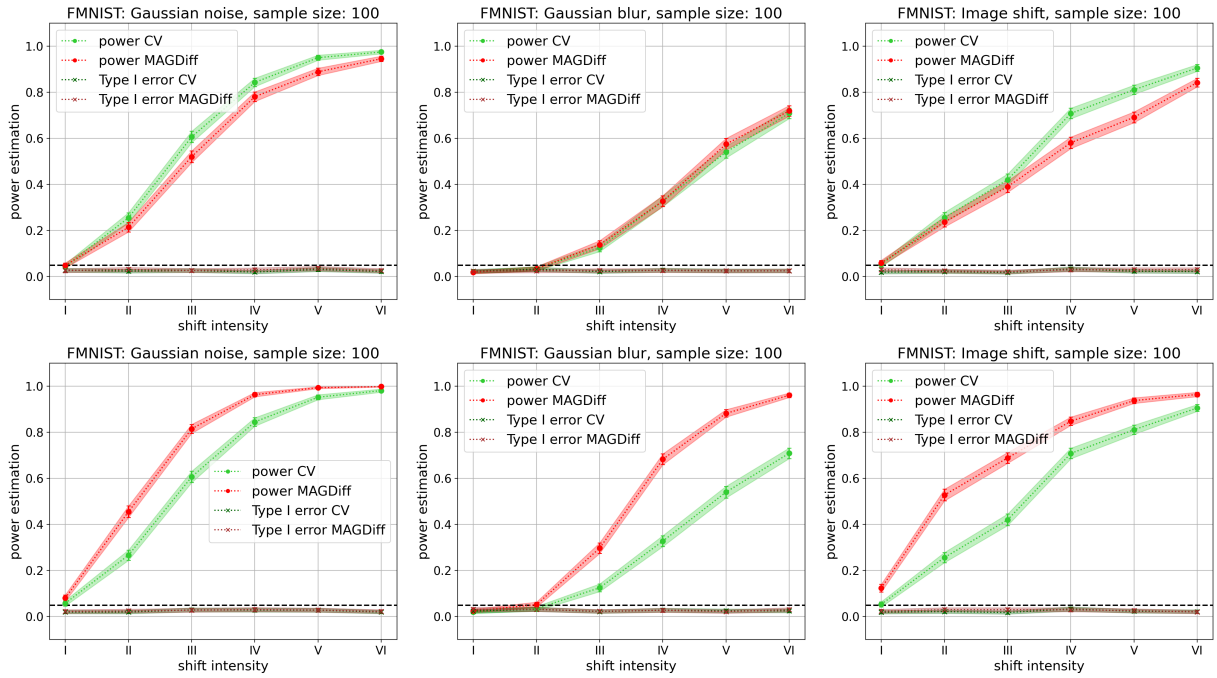

Figure 8: Power and type I error of the test with `MAGDiff` (red) and CV (green) representations w.r.t. the shift intensity for various shift types on the FMNIST dataset with $\delta = 0.5$, sample size 100, for layers ℓ_{-1} (top row) and ℓ_{-3} (bottom row).

A.3 Theoretical observations regarding the preservation of shift distributions by continuous functions

In the main article, we mentioned the fact that under generic conditions, two distinct distributions remain distinct under the application of a non-constant continuous function (though this does not necessarily translate to good quantitative guarantees). In this section, we make this assertion more formal and provide an elementary proof.

Let X be a separable metric space, and denote by $\mathcal{P}(X)$ the set of probability measures on X equipped with its Borel σ-algebra. Let $C_b(X)$ be the real bounded continuous functions on X. We consider the weak convergence topology on $\mathcal{P}(X)$; remember that a subbase for this topology is given by the sets

$$U_{f,a,b} := \left\{ \mu \in \mathcal{P}(X) | \int_X f d\mu \in]a, b[\right\},$$

for $f \in C_b(X)$ and $a < b \in \mathbb{R}$ (see for example Kallianpur (1961)).

Now let X, Y be two such separable metric spaces with their Borel σ-algebra. Any measurable map $F : X \to Y$ induces a map

$$\begin{aligned} F_* : &\mathcal{P}(X) \to \mathcal{P}(Y) \\ &\mu \mapsto F_*(\mu), \end{aligned}$$

where $F_*(\mu)$ is the pushforward of μ by F, that is the measure on $\mathcal{P}(Y)$ characterized by $F_*(\mu)(A) = \mu(F^{-1}(A))$ for any Borel set $A \subset Y$.

Fact 1. *If $F : X \to Y$ is continuous, then $F_* : \mathcal{P}(X) \to \mathcal{P}(Y)$ is continuous for the weak convergence topology.*

Proof. Given $f \in C_b(X)$ and $a < b \in \mathbb{R}$, we see that $F_*^{-1}(U_{f,a,b}) = U_{f \circ F,a,b}$, which is enough to conclude by the definition of subbases. □

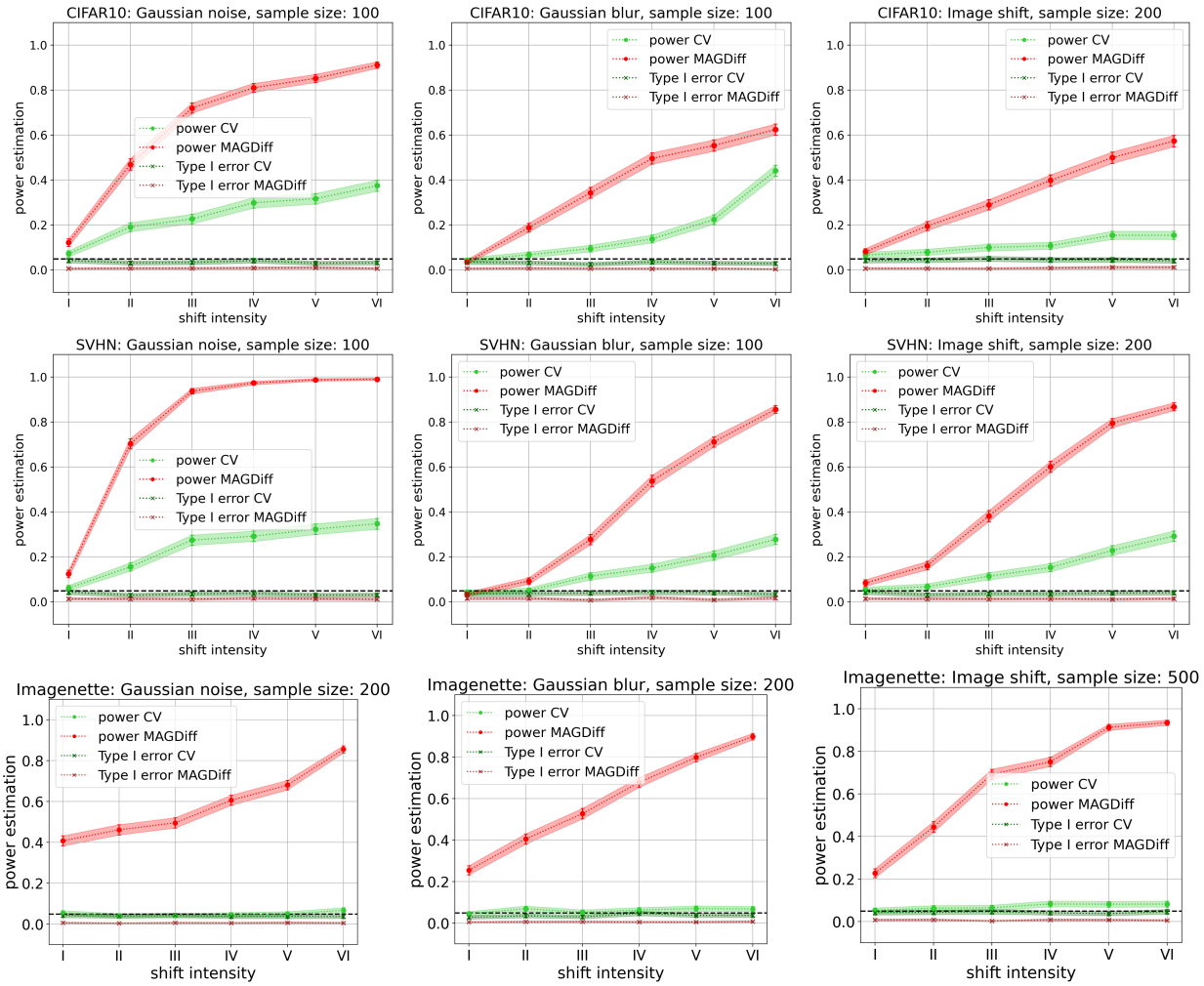

Figure 9: Power and type I error of the test with `MAGDiff` (red) and CV (green) representations w.r.t. the shift intensity for various shift types on the CIFAR-10, SVHN (with $\delta = 0.5$) and Imagenette (with $\delta = 1$) datasets. Sample sizes and values of δ were chosen to make the plots as expressive as possible (low power for low shift intensity, high power for high shift intensity), as the difficulty of the task varies depending on the shift type and dataset.

The following result follows from standard arguments; we give an elementary proof for the convenience of the reader.

Proposition 1. *Let $F : X \to \mathbb{R}$ be continuous and non-constant for X a separable metric space, and let $\nu \in F_*(\mathcal{P}(X)) \subset \mathcal{P}(\mathbb{R})$. Then the complement $F_*^{-1}(\{\nu\})^c = \mathcal{P}(X) \backslash F_*^{-1}(\{\nu\})$ of the set $F_*^{-1}(\{\nu\})$ is a dense open set of $\mathcal{P}(X)$ for the weak topology.*

Proof. As $\mathbb{R}$ is separable and metric, it is easy to show that the singleton $\{\nu\} \in \mathcal{P}(\mathbb{R})$ is closed (see for example (Kallianpur, 1961, Thm 4.1)). As we know from Fact 1 that F_* is continuous, we conclude that $F_*^{-1}(\{\nu\})$ is closed and $F_*^{-1}(\{\nu\})^c$ is open.

It remains to show that it is dense in $\mathcal{P}(X)$. Let μ belong to $F_*^{-1}(\{\nu\})$, and let $V \subset \mathcal{P}(X)$ be an open set containing μ. We have to show that $F_*^{-1}(\{\nu\})^c \cap V$ is non-empty. Thanks to the definition of the weak topology, we can assume (by potentially taking a subset of V) that

$$V = \bigcap_{i=1}^{n} \left\{ \tilde{\mu} \in \mathcal{P}(X) \text{ s.t. } \int_X f_i d\tilde{\mu} \in]a_i, b_i[\right\}$$

for some $f_1, \ldots, f_n \in C_b(X)$ and $a_1, b_1, \ldots, a_n, b_n \in \mathbb{R}$ with $a_i < b_i$ for all i. Let x_1 be any point in the support of μ. Then $\mu(B(x_1, \delta)) > 0$ for all $\delta > 0$ by definition of the support. As F is non-constant, there exists $x_2 \in X$ such that $F(x_2)$ is not equal to $F(x_1)$. Let us assume that $F(x_1) > F(x_2)$ (the proof is similar if $F(x_2) > F(x_1)$). By continuity, there exists $\epsilon > 0$ such that $F(x) > F(x_2)$ for any $x \in B(x_1, \epsilon)$. Define $m := \mu(B(x_1, \epsilon)) > 0$. For $t \in]0, 1[$, we define a new measure μ_t as follows : for any measurable set A, we let

$$\mu_t(A) = \mu(A \backslash B(x_1, \epsilon)) + (1-t)\mu(B(x_1, \epsilon) \cap A) + tm1_{x_2 \in A}.$$

For any such $t \in]0, 1[$, observe that

$$\begin{aligned} F_*(\mu_t)(]F(x_2), +\infty[) &= \mu_t(F^{-1}(]F(x_2), +\infty[) \\ &= F_*(\mu)(]F(x_2), +\infty[) - t\mu(B(x_1, \epsilon)) \\ &< F_*(\mu)(]F(x_2), +\infty[), \end{aligned}$$

which shows that $F_*(\mu) \neq F_*(\mu_t)$, hence that $\mu_t \in F_*^{-1}(\{\nu\})^c$.

On the other hand, we see that $|\int_X f_i d\mu_t - \int_X f_i d\mu| < 2tm||f_i||_\infty$ for $i = 1, \ldots, n$. Since $\mu \in V = \bigcap_{i=1}^n \{\tilde{\mu} \in \mathcal{P}(X) \text{ s.t. } \int_X f_i d\tilde{\mu} \in]a_i, b_i[\}$, thus $\mu_t \in V$ for $t \in]0, 1[$ small enough. This shows that $V \cap F_*^{-1}(\{\nu\})^c$ is non-empty, and thus we conclude that $F_*^{-1}(\{\nu\})^c$ is dense in $\mathcal{P}(X)$.

□

As a direct corollary, we get the following statement, where *generic*, as above, means that the property is true for any random variable x' whose distribution belongs to a fixed dense open set of the space of distributions on $\mathbb{R}^n$:

Corollary 1. *Let $F : \mathbb{R}^n \to \mathbb{R}^k$ be a non-constant continuous function represented by a neural network, and let x be a random variable on $\mathbb{R}^n$. For a generic random variable x' on $\mathbb{R}^n$, the distribution of $F(x')$ will be different from that of $F(x)$.*

Proof. $\mathbb{R}^n$ is a separable metric space, and if F is non-constant, so is at least one of its coordinate functions $F_i : \mathbb{R}^n \to \mathbb{R}$, to which Proposition 1 then applies. If the distribution of $F_i(x')$ is different from that of $F_i(x)$, then the distribution of $F(x')$ is different from that of $F(x)$. □

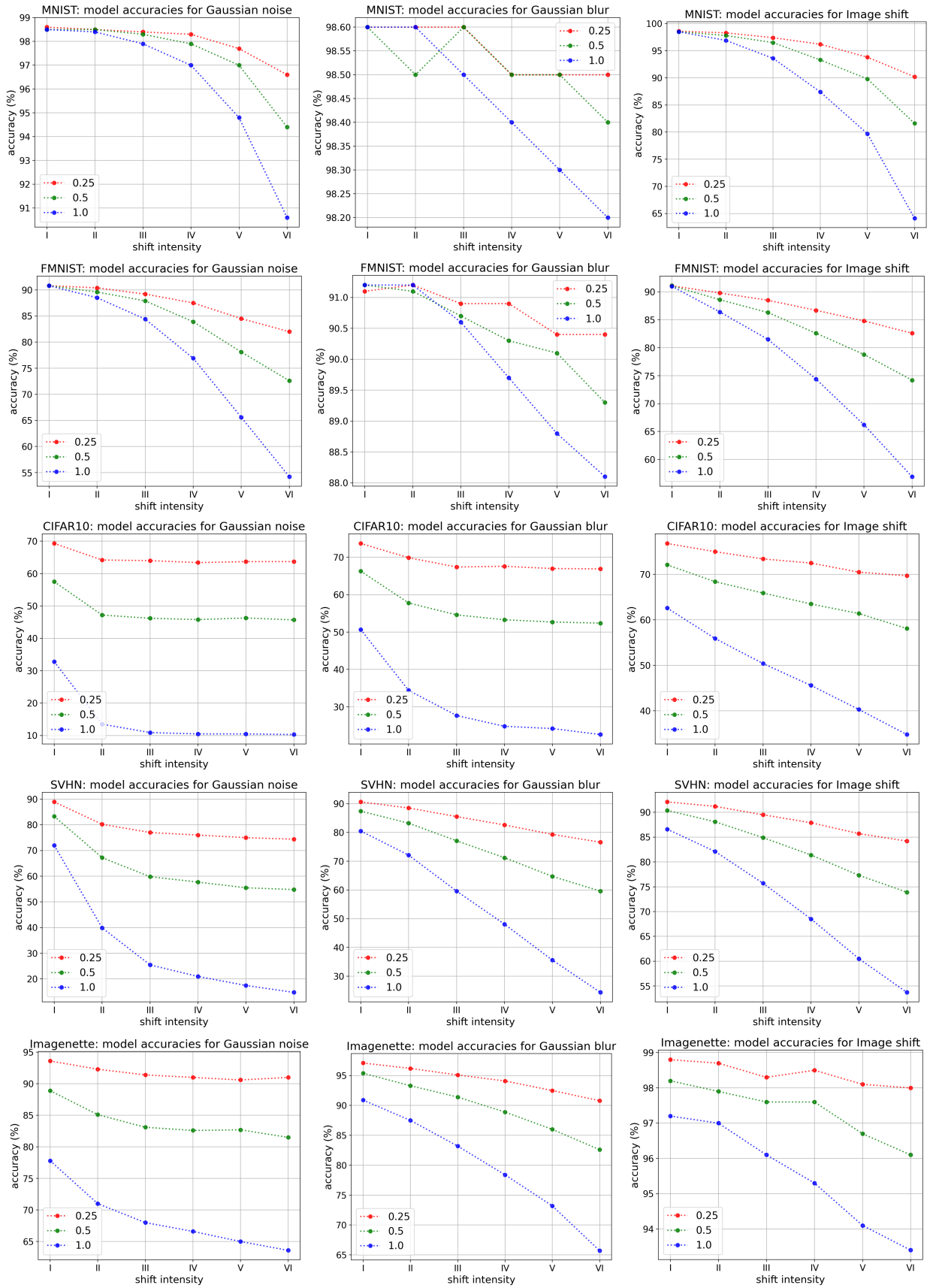

Figure 10: The impact of the shift type and intensity on the model accuracy for $\delta = 1.0$ (blue), $\delta = 0.5$ (green) and $\delta = 0.25$ (red).

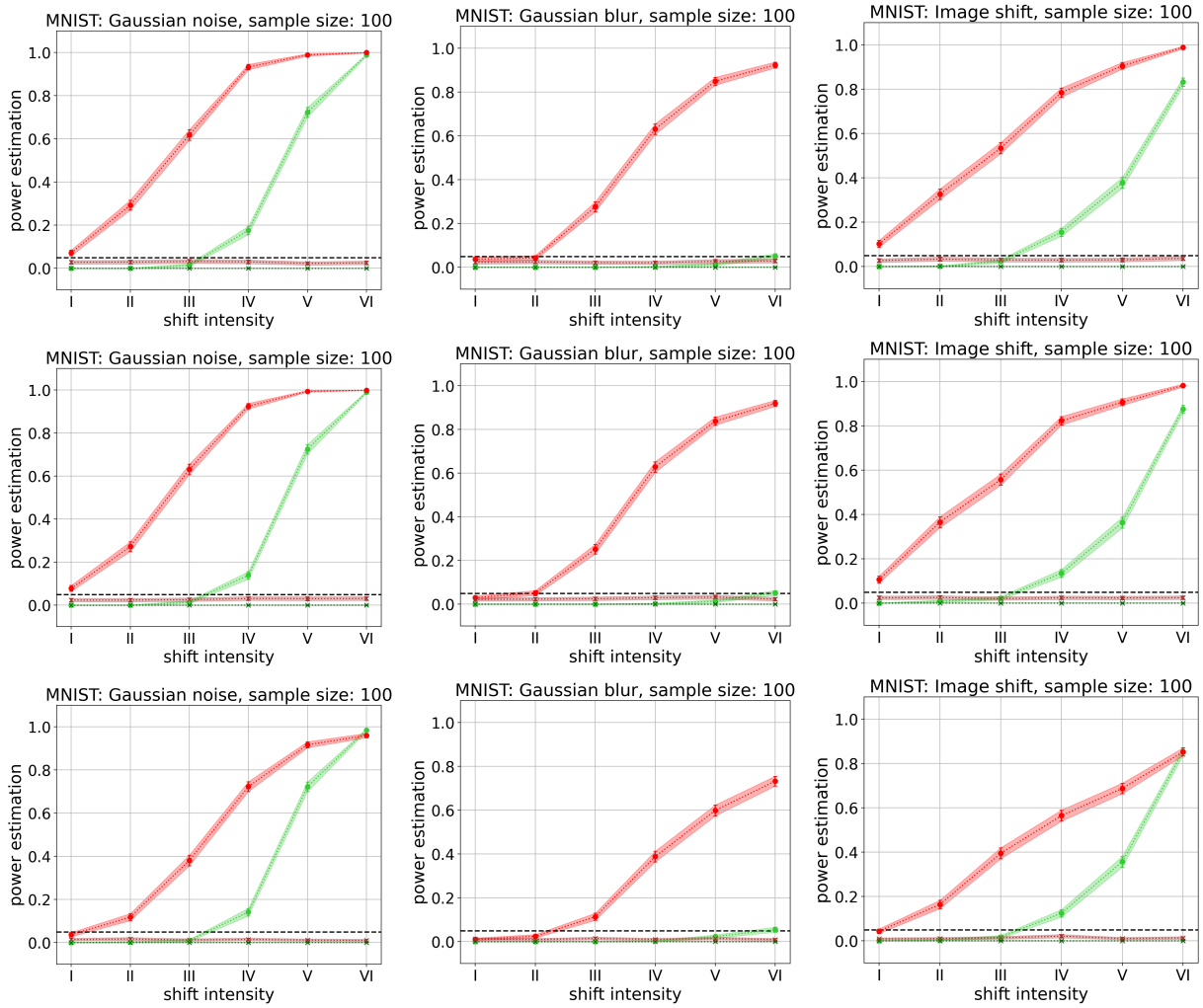

Figure 11: Power and type I error of the test with `MAGDiff` (red) w.r.t. the Frobenius norm, used in all other experiments, (top row), the spectral-norm (middle row) and $\|\cdot\|_\infty$ (bottom row) and CV (green) representations w.r.t. the shift intensity for various shift types on the MNIST dataset with $\delta = 0.5$, sample size 100, for layer ℓ_{-1}.