

---

# A study of data and label shift in the LIME framework

---

**Amir Hossein Akhavan Rahnama**

Department of Computer Science  
KTH Royal Institute of Technology  
Stockholm, Sweden  
arahnama@kth.se

**Henrik Boström**

Department of Computer Science  
KTH Royal Institute of Technology  
Stockholm, Sweden  
hbostrom@kth.se

## Abstract

LIME is a popular approach for explaining a black-box prediction through an interpretable model that is trained on instances in the vicinity of the predicted instance. To generate these instances, LIME randomly selects a subset of the non-zero features of the predicted instance. After that, the perturbed instances are fed into the black-box model to obtain labels for these, which are then used for training the interpretable model. In this study, we present a systematic evaluation of the interpretable models that are output by LIME on the two use-cases that were considered in the original paper introducing the approach; text classification and object detection. The investigation shows that the perturbation and labeling phases result in both data and label shift. In addition, we study the correlation between the shift and the fidelity of the interpretable model and show that in certain cases the shift negatively correlates with the fidelity. Based on these findings, it is argued that there is a need for a new sampling approach that mitigates the shift in the LIME’s framework.

## 1 Introduction

Local surrogate methods go back a long way in the literature of interpretable machine learning, see e.g. [Schmitz et al., 1999]. Local Interpretable Model-agnostic Explanations (LIME) is a recent example of these post-hoc methods, which has received significant attention [Ribeiro et al., 2016]. LIME provides an explanation for a single instance using a local surrogate, where the black-box is trained on the original input space and the surrogate is trained on an interpretable space. LIME’s usefulness lies in its flexibility to provide explanations for different data types, like text and images, while being model-agnostic.

In the next section, we briefly discuss some related studies. In Section 3, we present a novel approach of measuring data and label shift in LIME<sup>1</sup>, and apply it to the original case studies considered in [Ribeiro et al., 2016], as described in Section 4. Finally, in Section 5, we summarize the main conclusions and point out some directions for future research.

---

<sup>1</sup>LIME offers two algorithms for interpretability: Sparse Local Linear Explanations and SP-LIME. For clarity, this study considers the former only.

## 2 Related work

In [Alvarez-Melis and Jaakkola, 2018], a wide range of explanation methods for image classification were investigated. The study focused on empirically investigating *robustness* of explainability methods, and it showed that even minor changes in an input image can cause LIME to produce different explanations with a substantial variance in the selected (interpretable) features. The importance of locality for obtaining models with high fidelity was highlighted by [Laugel et al., 2018]. A method called **LocalSurrogate** was proposed to improve the sampling procedure of LIME. Some experiments on a set of UCI datasets were performed to show the usefulness of the approach. However, this approach still employs random perturbation, and the consequences of this choice will be investigated in this work. In Zhang et al. [2019], the authors showed that there is a significant level of uncertainty present even in LIME’s explanations of black-boxes with high test accuracy on synthetic and some UCI repository datasets.

## 3 Method

A standard assumption in machine learning is that training and test data are coming from the same distribution. One approach to detect if the assumption is violated is to calculate the divergence between the two samples and decide whether a shift has occurred, see [Quionero-Candela et al., 2009]. The divergence between the two samples can be measured in many ways, e.g., using KL-divergence or Bregman divergence. Maximum mean discrepancy [Gretton et al., 2012] is chosen here, since it not only allows for being computed reasonably fast, i.e., in quadratic time, but also provides an acceptance threshold that can be computed without extra constraints on the distributions being tested. In addition to that, it comes with a two sample testing framework with theoretical guarantees, see [Gretton et al., 2012] for more details.

Assume that we are interested in the explanation of an instance  $x$  that is predicted by a black-box model  $f$ , for which we get a numerical score, e.g., an estimate of the class probability, with respect to some specific class label  $y$ , here denoted by  $f_y(x)$ , together with a corresponding score for the class label output by the local surrogate, here denoted  $g_y(x)$ . We then define the fidelity of  $g$  to  $f$  as follows:

$$\mathbb{F}(x, y) = \frac{1}{|f_y(x) - g_y(x)| + 1} \quad (1)$$

In this work, the above formula is used for calculating the fidelity between an interpretable model and the underlying black-box model with respect to a specific set of instances.

## 4 Empirical investigation

In this section, we aim to answer the following questions about the LIME framework:

- Does the perturbed instances ( $Z$ ) come from another distribution than the original training instances ( $X_{\text{train}}$ ), i.e., has a data shift occurred?
- If the above holds, do the black-box predictions for the two sets ( $\{f(z) : z \in Z\}$  and  $\{f(x) : x \in X_{\text{train}}\}$ ) come from different distributions, i.e., has a label shift occurred?
- Does the above shifts have a negative effect on the fidelity of the interpretable model?

### 4.1 Text classification with SVM

In this experiment, we provide explanations for all the test instances in the Newsgroups dataset with regards to the class *atheism*. The aim is to show both the magnitude of divergence and frequency in which shift occurs. To investigate the local neighborhood of an explained instance  $x$ ,  $n$  neighbouring training instances (using the cosine kernel) are selected, here denoted as  $X_{\text{knn}}$ . After that, for our data shift test, we run a two sample MMD kernel test ( $\alpha = 0.05$ ) between  $X_{\text{knn}}$  and the perturbed instances ( $Z$ ). In this case, the null hypothesis is  $H_0 : P_{X_{\text{knn}}} = P_Z$ . For the label shift test, we run a two sample MMD kernel test ( $\alpha = 0.05$ ) between  $F_{X_{\text{knn}}} = \{f(x) : x \in X_{\text{knn}}\}$  and  $F_Z = \{f(z) : z \in Z\}$ . In this case, the null hypothesis is  $H_0 : P_{F_{X_{\text{knn}}}} = P_{F_Z}$ .

Table 1: Data shift two sample test results for Newsgroup test instances:  $H_0 : P_{X_{\text{knn}}} = P_Z$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
2	417 (57%)	300 (43%)	$0.42 \pm 0.34$
20	717 (100%)	0 (0%)	$5.56 \pm 1.58$
100	717 (100%)	0 (0%)	$24.77 \pm 8.00$
200	717 (100%)	0 (0%)	$44.20 \pm 15.84$
500	717 (100%)	0 (0%)	$87.35 \pm 36.75$

Table 2: Label shift two sample test results for Newsgroup test instances along with fidelity measures:  $H_0 : P_{F(X_{\text{knn}})} = P_{F(Z)}$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
2	239 (33.3%)	478 (66.6%)	$0.21 \pm 0.56$
20	515 (71.8%)	202 (28.1%)	$2.38 \pm 1.93$
100	716 (99.8%)	1 (0.2%)	$11.97 \pm 7.69$
200	717 (100%)	0 (0%)	$24.29 \pm 14.47$
500	717 (100%)	0 (0%)	$63.06 \pm 31.16$

In Table 1, the frequency of accepted tests with different values of  $n$  are shown. As can be seen, as the value of  $n$  increases, the divergence values increases significantly. Table 1 shows that for sample sizes that are larger than 20, the first null hypothesis stated above can be rejected for *all test instances*. The results of the tests for label shift are presented in Table 2. Similar to the previous test, as the number of samples,  $n$ , increases, the divergence becomes larger. The increase is at a lower pace compared to the test for the data shift in this case, nonetheless with even  $n = 2$ , more than a third of the two sample test cases can be rejected. The experiments have hence shown that the random perturbation of LIME may result in considerable data and label shift even for small sample sizes.

In Figure 1, the average fidelity of the interpretable model output by LIME over all test instances is displayed for varying sample sizes. Although, there is no widely accepted fidelity criteria for surrogate models like LIME, the achieved fidelity rates can hardly be accepted as they are not much better than random.

## 4.2 Object detection with deep neural networks

In this use-case, LIME explanations are calculated for the top-1 predicted class label of each test instance test instances in ImageNet ([Deng et al., 2009]). Due to the fact that it is computationally infeasible to find nearby instances via K-Nearest Neighbours directly in the ImageNet training dataset, an alternative approach is considered here: for each explained instance ( $x$ ), a random sub-sample from the images with the class equal to the predicted class of the black-box model, namely  $F(x)$  is compared against the perturbed samples **LIME**, namely  $Z$ . This random sub-sample is called  $X_{\text{local}}$  and since  $F$  is an accurate black-box model trained on  $X_{\text{train}}$  and  $x \in X_{\text{train}}$ , therefore our assumption is that the predicted value of  $F(x)$  can help to find similar instances to  $x$  in  $X_{\text{train}}$  without the need for knowing the ground truth with only negligible error. The test layouts for both data and label shift are exactly equal to those in 4.1, if the notation of  $X_{\text{knn}}$  is replaced with  $X_{\text{local}}$ .

In this use-case, due to our limited computational budget, we have performed the tests on a sub-sample of 200 instances in **Imagenet**<sup>2</sup>. In Table 7 and Table 8, the results of our experiments are shown. In both experiments, the mean divergence of data and label shift and the number of rejected two sample tests are larger when compared to the the former use-case. As before, even with small number of perturbations, significant data and label shifts are visible. In Figure 1, it can be seen that in ImageNet example, the fidelity of the interpretable model is worse than the Newsgroup use-case on average.

Due to the overall low fidelity in the ImageNet use-case study, one may argue that the quality of the output explanations of LIME for object detection using deep neural networks can be questioned.

<sup>2</sup>See additional material for the predicted class distribution of this sample

Table 3: Data shift two sample test results for ImageNet test instances for their top predicted class label:  $H_0 : P_{X_{\text{local}}} = P_Z$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
50	188 (100 %)	188 (0%)	$6.56 \pm 0.13$
100	188 (100%)	0 (0%)	$13.16 \pm 0.20$
200	188 (100%)	0 (0%)	$26.21 \pm 0.35$
500	188 (100%)	0 (0%)	$65.32 \pm 0.67$

Table 4: Label shift two sample test results for ImageNet test instances for their top predicted class label:  $H_0 : P_{F(X_{\text{local}})} = P_{F(Z)}$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
50	188 (100 %)	0 (0%)	$34.18 \pm 6.13$
100	188 (100%)	0 (0%)	$69.03 \pm 12.73$
200	188 (100%)	0 (0%)	$139.16 \pm 25.63$
500	188 (100%)	0 (0%)	$346.30 \pm 67.99$

## 5 Conclusion

In this study, we have presented experimental results showing that instances generated by LIME’s perturbation procedure are significantly different from training instances drawn from the underlying distribution. Our method of choice for detecting the shift is the MMD kernel two sample kernel test. The experimental results also investigated the correlations of the fidelity of interpretable model obtained by LIME with the shift. In some cases, the shift can have a negative correlation on the fidelity of the interpretable model. Based on the results from the tests, we argue that random perturbation of features of the explained instance cannot be considered a reliable method of data generation in the LIME framework.

In order to mitigate these shifts, alternative approaches to the perturbation method in LIME may be considered. One natural alternative would be to consider sampling from training instances in the vicinity of the explained instance rather than perturbing the features of the instance.

## Acknowledgment

Amir Hossein Akhavan Rahnama would like to thank Patrik Tran and Amir Payberah for their valuable feedback in the process of writing this paper.

## References

Gregor PJ Schmitz, Chris Aldrich, and Francois S Gouws. Ann-dt: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10(6):

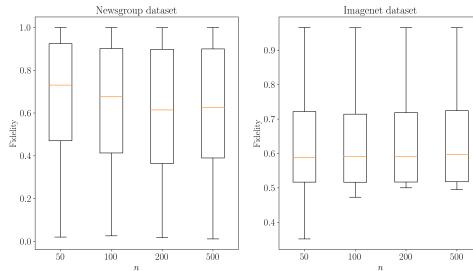


Figure 1: Fidelity and MMD divergence in Newsgroup and ImageNet dataset

1392–1401, 1999.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretability. *Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. “why should you trust my explanation?” understanding uncertainty in lime explanations. *ICML2019 Workshop AI for Social Good*, 2019.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 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(Mar):723–773, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

## A. Interpretable Machine Learning

Many state-of-the-art machine learning models are essentially black-boxes, in the sense that they have a highly nonlinear internal structure that makes a global understanding of their predictions hardly possible in practice. Post-hoc interpretable methods are used to explain the predictions of such black-box models after they are trained. A subclass of post-hoc interpretable methods is called surrogate methods, by which the black-box model is approximated with another (interpretable) model to provide a better understanding of the former. A recent example of these types of method is the popular LIME approach, which is the focus of this work.

## B. Shift in Machine Learning

A standard assumption in machine learning is that training and test data are coming from the same distribution, namely  $P_{\text{train}}(x, y) \stackrel{D}{=} P_{\text{test}}(x, y)$ , where  $P_{\text{train}}(x, y)$  and  $P_{\text{test}}(x, y)$  denote the joint probability of an instance  $x$  and a label  $y$ . Since  $P(x, y) = P(y|x)P(x)$ , any inequality between the two joint distributions may come either from a shift of the prior (covariate shift), i.e.,  $P_{\text{train}}(x) \stackrel{D}{\neq} P_{\text{test}}(x)$ , the conditional, i.e.,  $P_{\text{train}}(y|x) \stackrel{D}{\neq} P_{\text{test}}(y|x)$  or both (concept drift). [Quionero-Candela et al., 2009].

## C. Maximum Mean Discrepancy

Assume that  $X$  and  $Y$  are random variables defined on a topological space  $\chi$ , with corresponding Borel probability measures  $p$  and  $q$ . Let us assume we have the independently and identically distributed observations  $X := \{x_1, \dots, x_m\}$  and  $Y := \{y_1, \dots, y_n\}$ <sup>3</sup>.

Let  $\mathcal{F}: \chi \rightarrow \mathcal{R}$  be a class of functions. MMD is defined as follows [Gretton et al., 2012]:

$$\text{MMD}_b[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} (\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]) \quad (2)$$

We extensively use the following two properties of MMD that are outlined in Theorem 1 and 2. For proof and a detailed description of the underlying assumptions, see [Gretton et al., 2012].

---

<sup>3</sup>This notation should not be confused with common usage of  $X$  and  $Y$  in supervised machine learning literature

---

**Algorithm 1** Sparse Linear Explanations using LIME

---

```

1: Input: Instance being explained  $x$ 
2: Input: Class label  $y$ 
3: Input: Black-box model  $f$ 
4: Input: Regularization path,  $\lambda$ 
5: Input: Number of samples,  $n$ 
6: Input: Number of features,  $K$ 
7: Input: Kernel function,  $\pi$ 
8: Output: Explanation,  $E$ 
9: Output: Loss function,  $L$ 
10: for  $i \leftarrow 1$  to  $n$  do
11:    $z_i \leftarrow$  Sample a random subset from non-zero features of  $x$  and zero out the rest of features
12:    $D_i \leftarrow \pi(z_i, x)$ 
13: end for
14:  $Z' \leftarrow$  Transform  $z_1, \dots, z_n$  into an interpretable representation
15:  $Z'_{\text{reg}} \leftarrow \lambda(Z')$  apply regularization path on  $Z'$ 
16: Train  $g$  using least squares with  $g(Z'_{\text{reg}})$  as inputs and  $f(Z)$  as labels and store the weights as
     $W_g$ 
17:  $E = \text{argmax}(|W_g|)_i$ s.t.  $i = 1, \dots, K$ 
18:  $L = D \times (f(z) - g(z'))^2$ 

```

---

**Theorem 1.** Let  $\mathcal{F}$  is a unit ball in a universal Reproducing kernel Hilbert space (RKHS) and is defined on a compact metric space and has a corresponding continuous kernel  $k(\cdot, \cdot)$ . Then  $MMD[\mathcal{F}, p, q] = 0$  if and only if  $p = q$ .

**Theorem 2.** A hypothesis test of level  $\alpha$  for the null hypothesis  $p = q$ , that is, for  $MMD[\mathcal{F}, p, q] = 0$ , has the acceptance region  $MMD_b[F, X, Y] < \sqrt{2K/m}(1 + \sqrt{2 \log \alpha^{-1}})$ .

Theorem 1 ensures that when MMD is zero, two distributions are equal. Theorem 2 provides an acceptable threshold to be used to reject the null hypothesis, i.e.  $p = q$ . It should be noted that this test requires no further assumption on the type or class of distributions for  $p$  and  $q$ .

## D. LIME

The LIME algorithm provides an explanation for an instance  $x$ ; see Algorithm 1). In addition to this instance, other inputs to the algorithm consists of the following: a trained black-box model  $f$ , a regularization path  $\lambda$ , the total number of perturbed instances  $n$ , the number of features in the explanation  $k$  and a kernel function  $\pi$ . LIME starts off by perturbing non-zero features  $x$  uniformly at random for  $n$  times. These new instances are called  $z_i (i = 1, \dots, n)$ . The corresponding distances of  $z_i$  (stored in the matrix  $Z$ ) from  $x$ , namely  $D$ , is calculated using the kernel function,  $\pi$ . After that, the  $Z$  is transformed into a binary representation called  $Z'$ . In the next step,  $Z'$  is then fed into the regularization path  $\lambda$  (stored in the matrix  $Z'_{\text{reg}}$ ) to reduce the dimensionality of co-linearly dependent features in  $Z$ . The interpretable model  $g$  is then trained using least squares on  $Z'_{\text{reg}}$  as inputs and the output of the black-box on  $Z$  with respect to class  $y$ , namely  $f_y(Z)$  as labels. After the interpretable model is fit, the weights are stored as  $|W_g|$ . At the end,  $K$  features of  $Z'_{\text{reg}}$  with the largest absolute weights in  $|W_g|$  are returned as explanations. LIME's loss function is used to measure the quality of explanations:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z')^2 \quad (3)$$

The loss is minimized when the output of the interpretable model on the interpretable representation,  $g(z'_{\text{reg}})$ , has the least difference from the output of the black-box model on  $Z$ , namely  $f(z)$ <sup>4</sup>.

---

<sup>4</sup>We can only speculate that the formula 3 is a variation of the fidelity of  $g$  with regards to  $f$ , as is not further explained nor evaluated in [Ribeiro et al., 2016]

Table 5: Data shift two sample test results for Newsgroup test instances:  $H_0 : P_{X_{\text{knn}}} = P_Z$  ( $\alpha = 0.05$ )

$T_n$	REJECT	FAILED TO REJECT	MMD
2	417 (57%)	300 (43%)	$0.42 \pm 0.34$
5	681 (94%)	36 (6%)	$1.34 \pm 0.52$
10	710 (99%)	7 (1%)	$2.82 \pm 0.88$
20	717 (100%)	0 (0%)	$5.56 \pm 1.58$
50	717 (100%)	0 (0%)	$13.19 \pm 4.03$
100	717 (100%)	0 (0%)	$24.77 \pm 8.00$
200	717 (100%)	0 (0%)	$44.20 \pm 15.84$
500	717 (100%)	0 (0%)	$87.35 \pm 36.75$

Table 6: Label shift two sample test results for Newsgroup test instances along with fidelity measures:  $H_0 : P_{F(X_{\text{knn}})} = P_{F(Z)}$  ( $\alpha = 0.05$ )

$T_n$	REJECT	FAILED TO REJECT	MMD
2	239 (33.3%)	478 (66.6%)	$0.21 \pm 0.56$
5	209 (29.1%)	508 (70%)	$0.62 \pm 0.78$
10	336 (46.8)	381 (1%)	$1.16 \pm 1.15$
20	515 (71.8%)	202 (28.1%)	$2.38 \pm 1.93$
50	697 (85.2%)	20 (0.02%)	$5.88 \pm 4.00$
100	716 (99.8%)	1 (0.2%)	$11.97 \pm 7.69$
200	717 (100%)	0 (0%)	$24.29 \pm 14.47$
500	717 (100%)	0 (0%)	$63.06 \pm 31.16$

## E. Details of experiments with additional figures and tables

### E1. Text classification with SVM

In this experiment, we have replicated the text classification use-case originally investigated in [Ribeiro et al., 2016]. The LIME approach is studied together with a black-model consisting of an SVM with a RBF kernel, and using ridge regression as the algorithm for generating surrogate models. The selected regularization path is Least Angle Regression LASSO, respectively and the kernel function is the cosine kernel. The task concerns binary classification of documents into *christianity* or *atheism*, where the documents come from the newsgroup dataset<sup>5</sup>, divided into 1079 training instances ( $X_{\text{train}}$ ) and 717 test instances( $X_{\text{test}}$ ). In this experiment, documents are transformed into the Term-Frequency (TF) representation, corresponding to a total of 19666 words. In this experiment, we limit the explanations to the class label *atheism*. The MMD divergence values for the data shift test, between the perturbed samples of LIME ( $Z$ ) and nearby instances of the explained instance  $X_{\text{knn}}$  can be seen in in Figure 2 and corresponding divergence values for the label shift divergence measures is shown in Figure 2. In addition, the detailed information on two sample tests between these values for different number of samples ( $n$ ) can be seen in Table 5. Lastly, a detailed overview of two sample tests between the predicted value of the black-box model on both samples, namely  $f(Z)$  and  $f(X_{\text{knn}})$  is shown in Table 6. The boxplot of the fidelity of the interpretable model for various sample size ( $n$ ) can be seen in Figure 4 along with mean fidelity values.

### E2. Object detecton with Deep Neural Networks

In this experiment, we have replicated the objected detection use-case originally investigated in LIME’s paper. The black-model is the pre-trained Inception V3 model, and ridge regression as the model for the surrogate. The selected regularization path is Least Angle Regression LASSO, and the kernel function is the cosine kernel. The dataset for this experiment is ImageNet dataset. The explained image is transformed into the super-pixels representation using Quicksift algorithm. The task concerns the multi-class object detection of images in the Imagenet dataset with 1001 classes. The dataset has 10000000 images in its training set, namely ( $X_{\text{train}}$ ) and 60000 in its test set, namely ( $X_{\text{test}}$ ). In our experiment, LIME explanations for the top predicted class are considered. In this use-case we have selected a sample of 200 instances from Imagenet (see Figure 5 for predicted class distribution of the instances).

The MMD divergence values for the data shift test, between the perturbed samples of LIME ( $Z$ ) and nearby instances of the explained instance  $X_{\text{local}}$  can be see in in Figure 6 and corresponding divergence values for the

<sup>5</sup><http://qwone.com/~jason/20Newsgroups/>

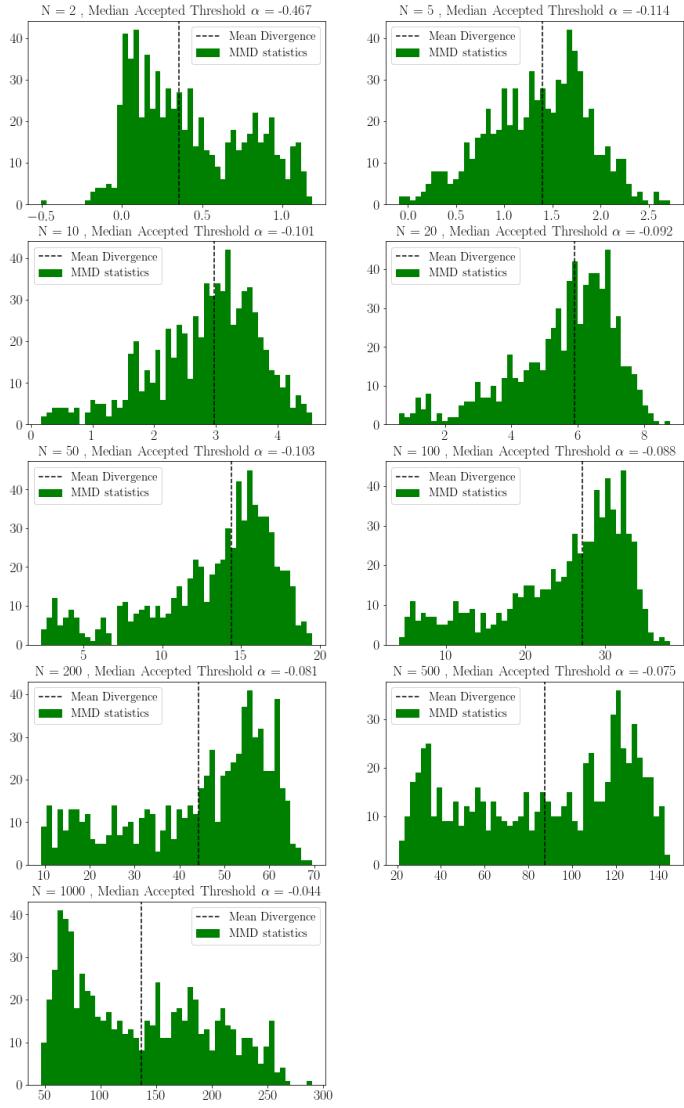


Figure 2: MMD Divergence values for data shift in the Newsgroup dataset

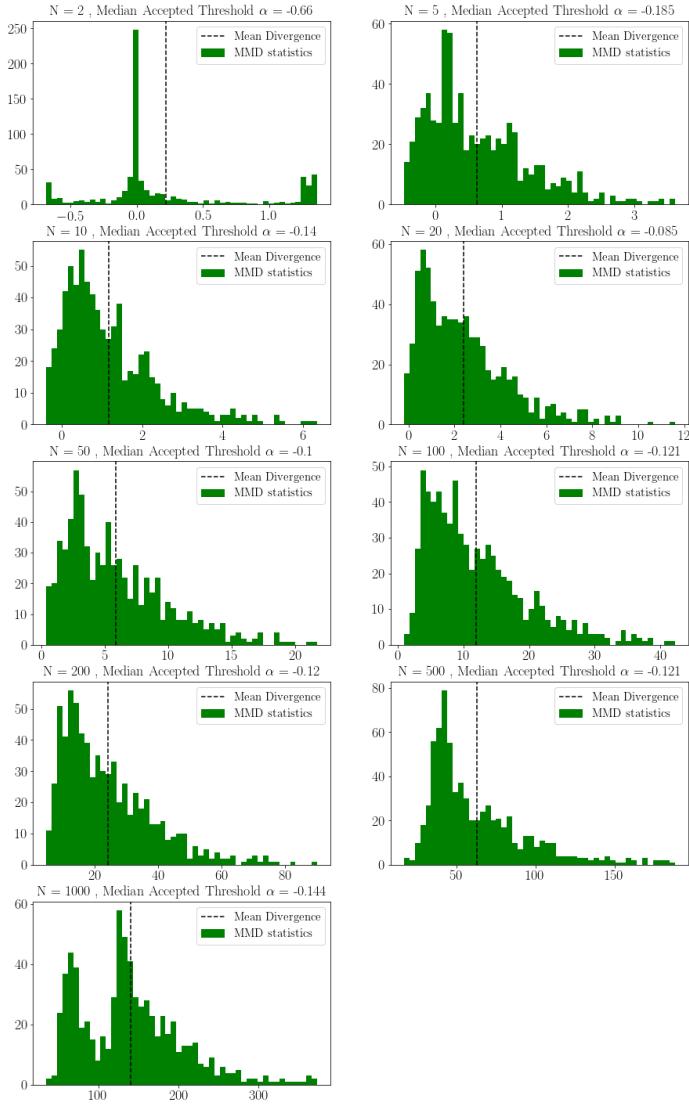


Figure 3: MMD Divergence values for label shift in the Imagenet dataset

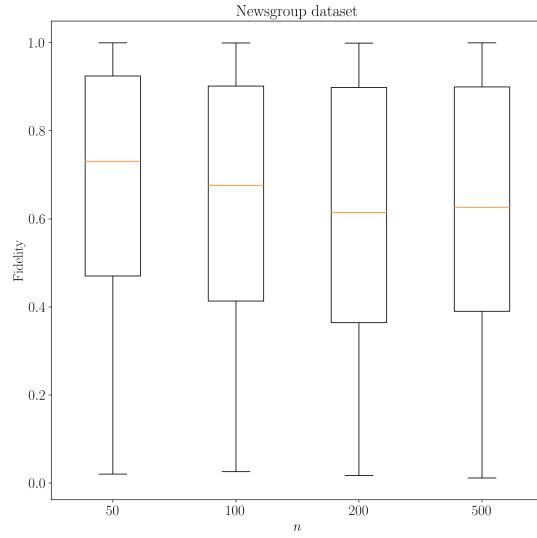


Figure 4: Fidelity of the LIME's interpretable model for explanations of atheism in Newsgroup dataset for each sample size  $n$

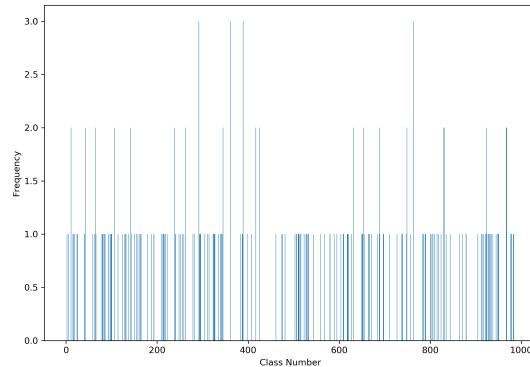


Figure 5: Predicted Class Frequency of our ImageNet sample by Inception V3

label shift divergence measures is shown in Figure 7. The boxplot of the fidelity of the interpretable model for various sample size ( $n$ ) can be seen in Figure 8 along with mean fidelity values. In Figure 9, the relationship between MMD divergence values and fidelity in both tests are shown.

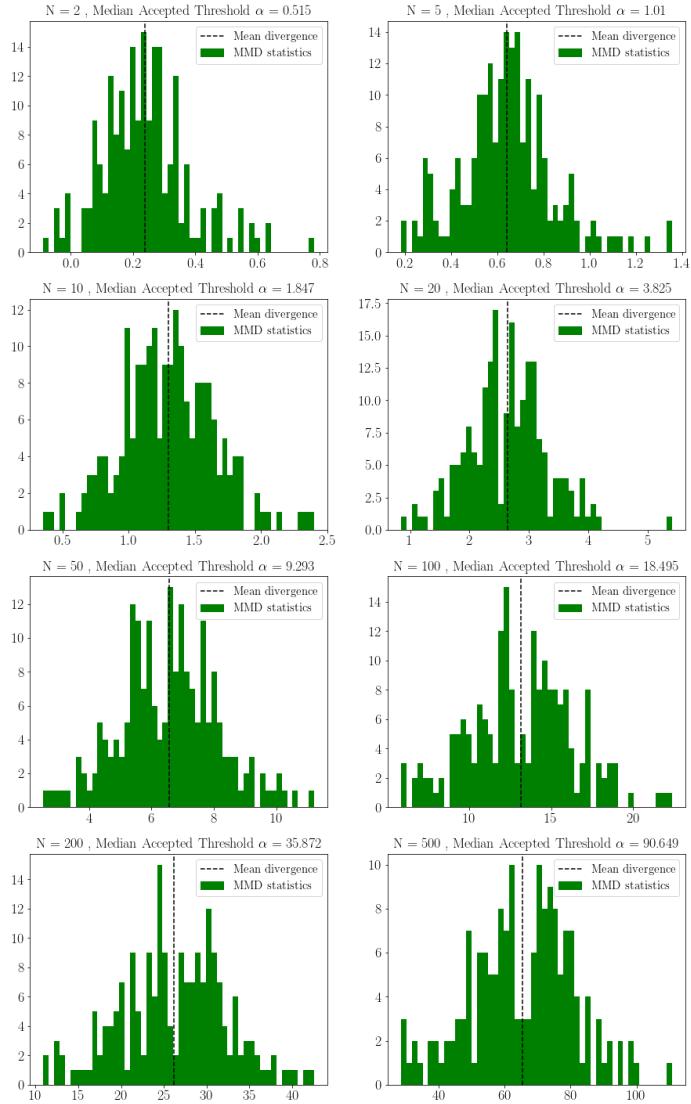


Figure 6: MMD Divergence values for data shift in the ImageNet dataset

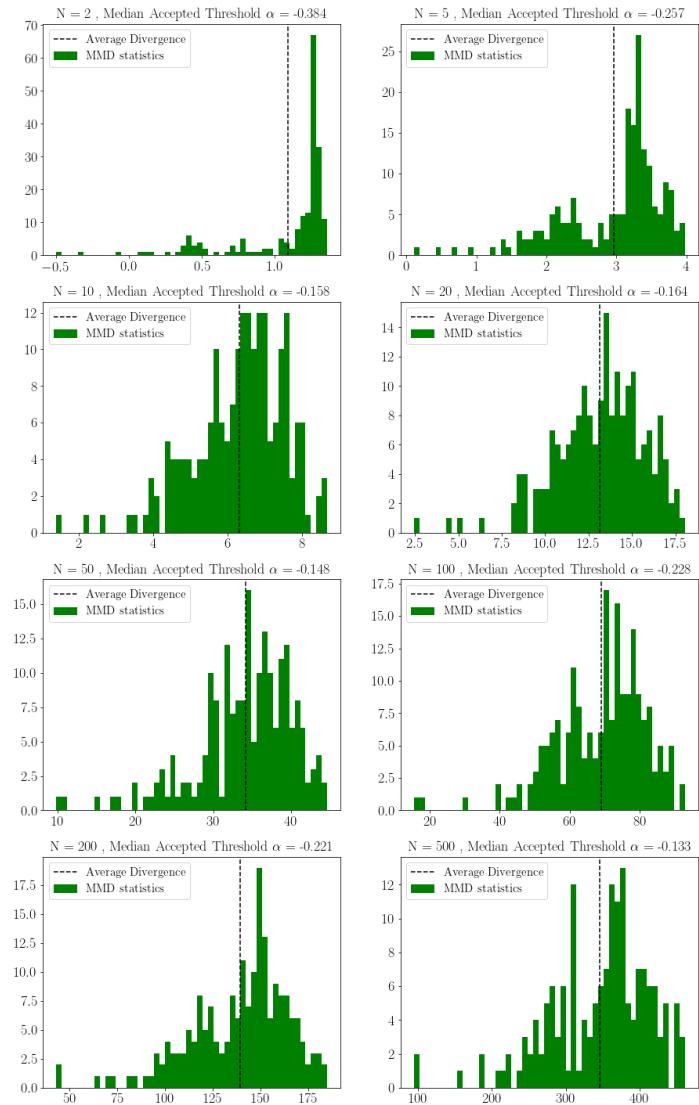


Figure 7: MMD Divergence values for label shift in the ImageNet dataset

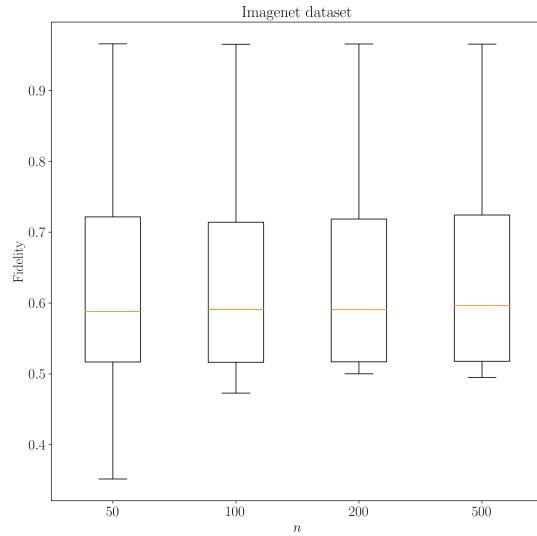


Figure 8: Fidelity histogram for LIME explanations of top-1 predicted class in ImageNet

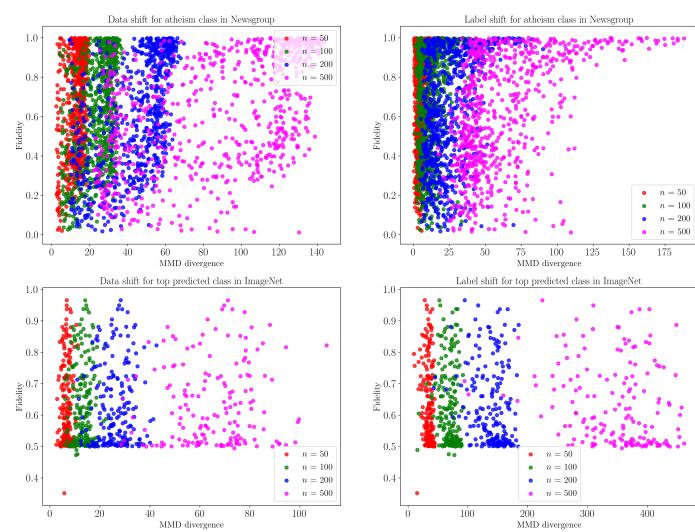


Figure 9: MMD divergence and fidelity in both use-cases

Table 7: Data shift two sample test results for ImageNet test instances for their top-1 predicted class label:  $H_0 : P_{X_{\text{local}}} = P_Z$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
2	86 (43 %)	113 (56%)	$0.23 \pm 0.13$
5	188 (100 %)	0 (0%)	$0.64 \pm 0.20$
10	188 (100 %)	0 (0%)	$1.29 \pm 0.35$
20	188 (100 %)	0 (0%)	$2.63 \pm 0.67$
50	188 (100 %)	0 (0%)	$6.56 \pm 0.13$
100	188 (100%)	0 (0%)	$13.16 \pm 0.20$
200	188 (100%)	0 (0%)	$26.21 \pm 0.35$
500	188 (100%)	0 (0%)	$65.32 \pm 0.67$

Table 8: Label shift two sample test results for ImageNet test instances for their top-1 predicted class label:  $H_0 : P_{F(X_{\text{local}})} = P_{F(Z)}$  ( $\alpha = 0.05$ )

$n$	REJECT	FAILED TO REJECT	MMD
2	188 (100 %)	0 (0%)	$1.08 \pm 0.34$
5	188 (100 %)	0 (0%)	$2.96 \pm 0.71$
10	188 (100 %)	0 (0%)	$6.31 \pm 1.23$
20	188 (100 %)	0 (0%)	$13.16 \pm 2.58$
50	188 (100 %)	0 (0%)	$34.18 \pm 6.13$
100	188 (100%)	0 (0%)	$69.03 \pm 12.73$
200	188 (100%)	0 (0%)	$139.16 \pm 25.63$
500	188 (100%)	0 (0%)	$346.30 \pm 67.99$