Detecting Adversarial Samples Using Density Ratio Estimates

Lovedeep Gondara
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
lgondara@sfu.ca

Abstract

Machine learning models, especially based on deep learning are used in everyday applications ranging from self driving cars to medical diagnostics. However, it is easy to trick such models using adversarial samples, indistinguishable from real samples to human eye, such samples can lead to incorrect classifications. Impact of adversarial samples is far-reaching and efficient detection of adversarial samples remains an open problem. In this paper we propose to use direct density ratio estimation as a model agnostic measure to detect adversarial samples, we empirically show that adversarial samples have different underlying probability densities compared to real samples. Our proposed method works well with colored and grayscale images, and with different adversarial sample generation methods.

1 Introduction

Adversarial samples are visually imperceptible perturbations to real data resulting in incorrect classifications with high confidence by machine learning models. At present, when machine learning models are deep embedded in our daily lives, existence of adversarial samples in real world poses a grave threat. Take an example of a self driving car, that uses a trained machine learning model to distinguish between different traffic signs, an adversarial sample can trick the model to classify stop sign as a yield sign as is shown by Papernot et al. [19], resulting in disastrous consequences. Figure 1 shows two such examples, including the traffic sign example, where models are tricked to incorrectly classify adversarially generated images that are otherwise indistinguishable to human eyes from original images. Crafting effective defenses against adversarial samples is an important research problem and an active research area. However, effective solutions remain elusive due to the limited understanding of the nature of adversarial samples.

In this paper we propose to use direct density ratio estimation [24, 25, 22] as an intuitive, simple and model agnostic approach to effectively distinguish adversarial samples from real samples. We empirically show that adversarially generated samples have a different underlying probability density compared to real samples, hence density ratio estimates are an effective tool to detect the differences and we show that density ratio estimation works for grayscale and colored images alike, and for different adversarial sample generation methods. We also study the effect of sample size on direct density ratio estimation to establish the effective sample size required to detect adversarial samples with high confidence.

Rest of the paper is organized as following: Section 2 presents preliminary introduction to direct density ratio estimation and methods for adversarial sample generation, section 3 presents insights to how and why density ratio estimation works for detecting adversarial samples. Section 4 presents empirical evaluation of our proposed method on MNIST and CIFAR-10 datasets followed by related work and our conclusions.
Figure 1: Example of adversarial input from Goodfellow et al. [5] ((a),(b),(c)) and Papernot et al. [19] ((d),(e)). (a) is the original target image correctly classified as “panda” by a model, (b) is the perturbation added, calculated using Fast Gradient Sign Method, and (c) is the resulting perturbed image, this added perturbation forces the classifier to output “gibbon” with 99.3% confidence. (d) is the real image of a stop sign, which is correctly classified as a stop sign, (e) is adversarially generated image, which is classified as a yield sign by the same model.

2 Preliminaries

In this section we introduce direct density ratio estimation method based on unconstrained least squares and methods for generating adversarial samples.

2.1 Density ratio estimation

Comparing probability distributions is a primary task in statistical learning, and has been used to study various problems [23, 25, 22]. One of the principled way of comparing two distributions \( p_a(x), p_b(x) \), from datasets \( a, b \), or the divergence between distributions is by estimating the density ratio

\[
r(x) = \frac{p_a(x)}{p_b(x)}
\]

(1)

A naive way of calculating \( r(x) \) would be to explicitly estimate \( p_a(x) \) and \( p_b(x) \), and plug the estimates in (1). But, direct density estimation is known to be a hard task [26].

This is easily overcome by directly estimating \( r(x) \), without having to estimate \( p_a(x) \) and \( p_b(x) \) separately [24, 17]. Different methods can be used to approximate \( r(x) \), but here we only focus on the efficient least squares approach [24].

\[
\hat{r}(x) = \alpha^T \phi(x)
\]

(2)

where \( \alpha = (\alpha_1, ..., \alpha_n)^T \) is the parameter to be learned and \( n \) is the number of parameters, \( \phi(x) = (\phi_1(x), ..., \phi_n(x))^T \) are the basis functions such that \( \phi(x) \geq 0_n \forall x \in D \), where \( D \) is the data domain and \( 0_n \) is the \( n \)-dimensional vector with all zeros. \( \alpha \) is determined to minimize the following squared error

\[
J_0(\alpha) = \frac{1}{2} \int \left( \frac{\hat{r}(x) - p_a(x)}{p_b(x)} \right)^2 p_b(x)dx = \frac{1}{2} \int \hat{r}(x)^2 p_b(x)dx - \int \hat{r}(x)p_a(x)dx + C
\]

(3)
where $C = \frac{1}{2} \int r(x)p_a(x)dx$ is a constant and can be ignored. Let,

$$J(\alpha) = J_0(\alpha) - C = \frac{1}{2} \alpha^T H \alpha - h^T \alpha$$  \hspace{1cm} (4)

where $H = \int \phi(x)\phi(x)^T p_b(x)dx$ and $h = \phi(x)p_a(x)dx$. Using empirical approximation, we obtain

$$\min_{\alpha \in \mathbb{R}^n} \left[ \frac{1}{2} \alpha^T \hat{H} \alpha - \hat{h}^T \alpha + \lambda_1 \alpha^T \alpha \right] \geq 0_n$$  \hspace{1cm} (5)

where $\hat{H} = \frac{1}{m_b} \sum_{i=1}^{m_b} \phi(x_i^b)\phi(x_i^b)^T$ and $\hat{h} = \frac{1}{m_a} \sum_{i=1}^{m_a} \phi(x_i^a)$. Here, $\lambda_1 \alpha^T \alpha$ is a regularization term, $\lambda \geq 0$, and $1_n$ is a $n$-dimensional vector of all ones.

An approximation of least squares approach, unconstrained least squares is given as

$$\min_{\beta \in \mathbb{R}^n} \left[ \frac{1}{2} \beta^T \hat{H} \beta - \hat{h}^T \beta + \frac{\lambda}{2} \beta^T \beta \right]$$  \hspace{1cm} (6)

(6) is obtained by ignoring the non-negativity constraints on the optimization problem given in (5). Advantage of using (6) is that the solution can be computed analytically. This approximation is the method used in this paper to estimate density ratio estimates. We encourage interested readers to read works of Kanamori and Hido \[7\] and Kanamori et al. \[8\] for further details of this method.

### 2.2 Generating adversarial images

Fast Gradient Sign Method (FGSM) for generating adversarial samples was introduced by Goodfellow et al. \[5\]. Given a model’s cost function $c(M, x, y)$, adversarial sample is generated as $x^* = x + \sigma_x$, where $\sigma_x$ is computed as

$$\sigma_x = \epsilon \text{sign}(\nabla_x c(M, x, y))$$  \hspace{1cm} (7)

where $\text{sign}(\nabla_x c(M, x, y))$ is the sign of model’s cost function gradient. $\epsilon$ controls the amount of perturbation, larger values of $\epsilon$ create highly perturbed images, distinguishable from real images by humans.

Jacobian Based Saliency Map Approach (JSMA) introduced by Papernot et al. \[20\] chooses perturbations by iteratively modifying only a limited number of features chosen based on decreasing adversarial saliency value where saliency values are calculated using model’s Jacobian matrix.

Target Gradient Sign Methods (TGSM) works to create perturbations in a way so as to push the misclassifications towards a specific class \[11\]. We refer readers to excellent works of Papernot et al. \[19\] and Kurakin et al. \[11\] for further details.

### 3 Density ratio estimation for detecting adversarial samples

This section shows why density ratio estimation is a good choice for detecting adversarial samples and how it is used in such scenarios.

#### 3.1 Why density ratio estimation works?

Implicitly, data are assumed to be generated from an underlying probability distribution with a certain probability density and a random sample of large size is often enough to represent the true data generating distribution. To create adversarial samples, original samples are perturbed in one way or the other. Perturbations, however small they are, result in perturbed density regions, that are different from original samples and easy to detect.

Figure 2 explicitly shows this phenomenon, where we have used tSNE \[16\] to plot original and FGSM created adversarial MNIST \[13\] and gray scaled CIFAR-10 \[10\] images from their respective
Figure 2: tSNE plots for MNIST and grayscaled CIFAR-10 test datasets and their adversarial counterparts generated using FGSM with $\epsilon = 0.3$, it is clear that densities of real and adversarial data differ considerably.

test partitions. We can clearly see the original images occupying certain density regions, which are very different when compared to their adversarially generated counterparts.

3.2 How density ratio estimation works?

We start with a simple example, where we only have real images, that is, our dataset does not have any adversarial images. We begin by drawing two random samples of sufficiently large size from the dataset. For now, sufficiently large can be assumed to be $n = 100$, and lets denote the two random samples by $X_1$ and $X_2$.

Using unconstrained least squared approach, described in section 2.1, we estimate the density ratio of $X_1$ and $X_2$ as

$$R(X) = \frac{p(X_1)}{p(X_2)}$$

where $p(X_1)$ and $p(X_2)$ are probability densities of $X_1$ and $X_2$ respectively. If $X_1$ and $X_2$ are from the same underlying probability distribution, we can see that $R(X)$ would be approximated to be closer to 1. And, if the samples are from different distributions, the ratio would be farther away from 1, as should be the case with adversarial samples, which indeed is true, as we do show in later sections.

3.3 What about colored images?

Until this section, we have only considered simple cases of grayscale images. Working with grayscale images is relatively easy, as we only have a single channel per image and density ratio estimation works fine. But, readers might wonder: What if the images are colored? as this is the case in most real life scenarios, and also, how does the adversarial perturbations affect colored images and their different channels.

Figure 3 shows a scenario of applying adversarial perturbations to colored images. Top row shows the tSNE plots of three color channels of original CIFAR-10 images from the test set and bottom
Figure 3: Density plots for colored CIFAR-10 images by color channel, generated using tSNE. Top row is the original images from the test set and the bottom row is the perturbed version for the same channel. Perturbation is clearly detectable in adversarial images.

row shows their respective perturbed versions created using FGSM with $\epsilon = 0.3$. It is clear that adversarially perturbed images have different densities compared to the original images, irrespective of the color channel. Hence, same density ratio estimation methods are applicable to colored images as are to grayscale images. In addition to the density ratio estimation of individual channels as defined in (8), we also define a density ratio estimate on the average of all three density ratio estimates of individual color channels, given as

$$R(X_a) = \frac{1}{3} \left( \frac{p(X^r_1)}{p(X^r_2)} + \frac{p(X^g_1)}{p(X^g_2)} + \frac{p(X^b_1)}{p(X^b_2)} \right)$$

where $X^r$, $X^g$, $X^b$ are red, green and blue channels respectively.

In following sections we show that either individual channels or the combined version can be used for detecting adversarial samples in colored images. The combined estimate can be used as a single statistic per comparison instead of using per channel individual estimates if required.

4 Evaluation

In this section, we present empirical evaluation of our proposed method on colored and grayscale images, with varying sample sizes and with varying adversarial sample generating methods.

4.1 Experimental setup

FGSM adversarial samples are generated using Cleverhans [18]. JSMA and TGSM samples are generated using Keras [2] and TensorFlow\footnote{Code adapted from: https://github.com/gongzhitaao/tensorflow-adversarial}. Datasets MNIST [12] and CIFAR-10 [9] are the primary datasets used in this study. CIFAR-10 is used twice, once in its original form and second time in its grayscale version. Test partitions of both datasets are used to generate adversarial samples. We define our setup for comparison using algorithm 1.

If adversarial samples are to be detected using density ratio estimation, we would expect $R_1$ to be very different from $R_2$, where $R_2$ would be closer to 1. For initial comparisons, we keep $m$ and $t$
Algorithm 1 Density ratio based adversarial sample detection

Require: $X$ and $Y$ as datasets with real and adversarial samples with $n$ samples in each.

for $i ← 1$ to $t$

    $a$: Sample a random index of length $m$ without replacement from $n$
    $b$: Sample a random index of length $m$ without replacement from $n$
    $x$: Sample from real data using index $a = X[\alpha,]$
    $y$: Sample from adversarial data using index $b = Y[\beta,]$
    $z$: Sample from real data using index $b = X[\beta,]$

    Estimate density ratio $R_1 = \frac{p(x)}{p(y)}$
    Estimate density ratio $R_2 = \frac{p(x)}{p(z)}$

end for

fixed at 100, that is, the experiments are run 100 times with sample size of 100 in numerator and denominator. Results are reported using mean $R_1$ and $R_2$ with related 95% confidence intervals.

4.2 Primary results

We start with the evaluation on grayscale images with adversarial samples generated using FGSM and varying values of $\epsilon$. Results are reported using average density ratios $R_1$ (Real-Adversarial) and $R_2$ (Real-Real) with their 95% confidence intervals.

Table 1 shows the results, it is clear that real-real density ratios are closer to 1, as we would expect them to be, compared to real-adversarial estimates. It is also seen that as value of $\epsilon$ increases, the density ratio for real-adversarial samples deteriorates dramatically. Statistical tests for comparing means can be used to easily generate hypothesis and produce p-values for comparisons.

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|------------------|-----------|
| MNIST      |                  |           |
| 0.1        | 1.83(1.75,1.91)  | 1.32(1.23,1.41) |
| 0.3        | 20.70(20.43,20.96) | 1.30(1.21,1.39) |
| 0.5        | 33.43(32.92,33.95) | 1.32(1.23,1.41) |
| 1          | 34.15(33.60,34.70) | 1.30(1.21,1.39) |
| CIFAR-Grayscale |              |           |
| 0.1        | 1.37(1,27,1,47)  | 1.02(0.99,1.04) |
| 0.3        | 30.21(29.43,30.99) | 0.99(0.99,1.0) |
| 0.5        | 52.03(50.28,53.78) | 1.00(0.99,1.0) |
| 1          | 51.28(49.8,52.7)  | 1.01(0.99,1.03) |

We have shown that density ratio estimation is a viable choice for detecting adversarial samples for grayscale images. For colored images, we use test partition of CIFAR-10 dataset and FGSM for adversarial sample generation, density ratio is estimated per color channel, that is, once each, for red, green and blue channels. We also estimate density ratio for combined channels, as described in section 3.3.

Table 2 shows the results of density ratio estimation on colored images. Similar to grayscale images, results show that real-real density ratio estimates are closer to 1, irrespective of $\epsilon$ values, whereas real-adversarial ratio estimates are significantly different, deteriorating with increasing $\epsilon$.

4.3 Varying sample size

Until now, we have compared the density ratio estimates using a fixed sample size of 100 for real and adversarial images. But in real life scenarios, we usually do not have 100 adversarial samples to estimate density ratio. So, to answer the question of how many samples are needed to reliably
Table 2: Density ratio estimates for colored images, comparisons are made per color channel and for combined channels, real-real density ratio is always closer to 1, compared to real-adversarial ratio estimates. Average Real-Adversarial density estimates for all values of $\epsilon$ are significantly different from Real-Real estimates using mean comparison with $\alpha \leq 0.05$.

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|------------------|-----------|
| 0.1        | 2.15(2.09,2.21)  | 0.99(0.99,1.0) |
| 0.3        | 30.28(29.58,30.99)| 1.00(0.98,1.02) |
| 0.5        | 41.73(40.49,42.97)| 1.02(0.99,1.05) |
| 1          | 41.76(40.43,43.0) | 1.00(0.99,1.00) |

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|------------------|-----------|
| 0.1        | 1.90(1.80,2.0)   | 1.00(0.99,1.0) |
| 0.3        | 28.52(27.85,29.20)| 1.00(0.99,1.00) |
| 0.5        | 45.65(44.27,47.04)| 1.00(0.99,1.01) |
| 1          | 45.68(44.35,47.00)| 1.00(0.99,1.01) |

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|------------------|-----------|
| 0.1        | 1.66(1.54,1.77)  | 1.00(0.98,1.03) |
| 0.3        | 19.07(17.90,20.25)| 1.01(0.98,1.05) |
| 0.5        | 38.0(36.86,39.13)| 1.04(1.0,1.09) |
| 1          | 36.59(35.48,37.70)| 1.03(0.99,1.06) |

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|------------------|-----------|
| 0.1        | 1.90(1.84,1.96)  | 1.00(0.99,1.01) |
| 0.3        | 25.96(25.21,26.71)| 1.00(0.99,1.02) |
| 0.5        | 41.79(40.99,42.59)| 1.02(1.00,1.04) |
| 1          | 41.34(40.52,42.16)| 1.00(0.99,1.02) |

Figure 4: Density ratio estimates with 95% confidence intervals by varying sample size for all three datasets, real-real density ratio is represented by blue line and real-adversarial density ratio estimates are represented by red line. It is seen that the density ratio estimates deteriorate with decreasing sample size, but a noticeable difference persists between real and adversarial estimates.

To estimate density ratios and to reject adversarial samples with confidence, we use all three datasets to run density ratio estimations, but with varying sample sizes from 80 to 10, both for real and adversarial samples, that is, we simultaneously decrease sample size for numerator and denominator when estimating density ratios.

Figure 4 shows the results, it is seen that density ratio estimates deteriorate as sample size decreases, there is a noticeable shift from the starting sample size of 80 going to 10. But, the difference between real and adversarial density ratio estimates persist, irrespective of number of samples. Using as few
as 10 samples, in numerator and denominator, it is possible to distinguish adversarial samples from real ones with high confidence.

### 4.4 Other attacks (JSMA and TGSM)

So far we have demonstrated the ability of density ratio estimation methods to distinguish between real and adversarial samples for FGSM generated samples. However, there are other methods of generating adversarial samples, how does density ratio estimation fare against samples that are not generated using FGSM?

To answer this question, we generate adversarial samples using JSMA and TGSM. We use a fixed sample size of 100 and the comparison strategy described in section 4.1.

Table 3: Density ratio estimates of real-adversarial and real-real samples using dataset MNIST and colored CIFAR-10, with adversarial samples generated using JSMA and TGSM methods, it is seen that density ratio estimates are capable of detecting adversarial samples irrespective of adversarial sample generation process. Average Real-Adversarial density estimates for all values are significantly different from Real-Real estimates using mean comparison with $\alpha \leq 0.05$.

| $\epsilon$ | Real-Adversarial | Real-Real |
|------------|-------------------|-----------|
| MNIST-JSMA | 2.34(2.28,2.40)   | 1.10(1.01,1.16) |
| CIFAR-10-red-JSMA | 2.83(2.63,3.04) | 1.00(0.98,1.02) |
| CIFAR-10-green-JSMA | 2.27(2.23,2.30) | 1.00(0.99,1.01) |
| CIFAR-10-blue-JSMA | 2.25(2.14,2.35) | 1.03(0.99,1.07) |
| CIFAR-10-comb-JSMA | 2.45(2.37,2.53) | 1.01(1.00,1.03) |
| MNIST-TGSM | 0.1   | 12.84(11.74,13.93) | 1.06(1.01,1.11) |
| MNIST-TGSM | 0.05  | 6.87(6.80,6.95)  | 1.09(1.04,1.15) |
| CIFAR-10-red-TGSM | 0.05  | 9.07(8.74,9.71)  | 1.03(0.99,1.06) |
| CIFAR-10-green-TGSM | 0.05  | 11.62(10.80,12.43) | 1.02(0.99,1.05) |
| CIFAR-10-blue-TGSM | 0.05  | 9.34(8.85,9.83)  | 1.03(0.99,1.07) |
| CIFAR-10-comb-TGSM | 0.05  | 10.00(9.65,10.37) | 1.03(1.01,1.05) |

Table 3 shows the results. It is seen that irrespective of adversarial sample generation process, adversarial samples have perturbed densities, hence, detection of adversarial samples with high confidence is possible using density ratio estimates.

### 5 Related work

Crafting effective defences against adversarial attacks is an active area of research. Several defences against adversarial attacks have been proposed, such as defensive distillation [21] and training models using adversarial examples [5], but they are generally computationally intensive and model specific, not model agnostic [19], similar to some other methods based on game theory [3, 15, 4]. Very recently, some interesting work has been done on adversarial sample detection, with Feinman et al. [4] working on detection of adversarial samples from artifacts, Li et al. [14] using outputs from convolutional layers to detect adversarial samples and Grosse et al. [6] investigating the statistical detection of adversarial examples. Our work is closer to [6] where we both propose model agnostic adversarial sample detection methods. However, our methods are primarily different, ours is based on direct density ratio estimation compared to maximum mean discrepancy in [6] and we provide detailed insight to densities of adversarial samples, and the utility of our method with colored images.

### 6 Conclusion

We have presented empirical evidence that adversarial samples have perturbed underlying probability densities compared to real samples and it is possible to detect adversarial samples with high confidence using density ratio estimates. We have also shown that density ratio estimates work well with colored and grayscale images and density ratio estimates are capable of detecting adversarial samples generated using varying methods with small sample sizes.
References

[1] M. Brückner and T. Scheffer. Stackelberg games for adversarial prediction problems. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 547–555. ACM, 2011.

[2] F. Chollet. Keras. https://github.com/fchollet/keras, 2017.

[3] N. Dalvi, P. Domingos, S. Sanghais, D. Verma, et al. Adversarial classification. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 99–108. ACM, 2004.

[4] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410, 2017.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[6] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280, 2017.

[7] T. Kanamori, S. Hido, and M. Sugiyama. Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection. In Advances in neural information processing systems, pages 809–816, 2009.

[8] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. Journal of Machine Learning Research, 10(Jul):1391–1445, 2009.

[9] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[10] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). 2009.

[11] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[13] Y. LeCun and C. Cortes. Mnist handwritten digit database. 2010.

[14] X. Li and F. Li. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767, 2016.

[15] W. Liu and S. Chawla. Mining adversarial patterns via regularized loss minimization. Machine learning, 81(1):69–83, 2010.

[16] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.

[17] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. 2007.

[18] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel. Cleverhans v1.0.0: an adversarial machine learning library. arXiv preprint arXiv:1610.00768, 2016.

[19] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697, 2016.

[20] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, pages 372–387. IEEE, 2016.

[21] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In Security and Privacy (SP), 2016 IEEE Symposium on, pages 582–597. IEEE, 2016.

[22] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000.

[23] A. J. Smola, L. Song, C. H. Teo, et al. Relative novelty detection. 2009.
Appendix

A: Working with small sample sizes, a real life perspective

In the paper we have talked about utility of density ratio estimates with varying sample sizes to detect adversarial samples. We demonstrated the capability of adversarial sample detection with decreasing sample size in numerator and the denominator of density ratio estimates.

In real life scenarios however, we always have enough supply of real images. Hence, here we investigate the effectiveness of using density ratio estimates to detect adversarial samples by only varying the sample size of adversarial images. As we have already shown the utility of our proposed method with sample size as small as 10, here we reduce the sample size for adversarial images even further, keeping the sample size for real images fixed at 100. As the results are similar on MNIST and CIFAR-10 datasets with different adversarial sample generation methods, here we only concentrate on samples generated using FGSM on MNIST data. We can define this new comparison as a modified version of comparison algorithm from paper, given as algorithm

\begin{algorithm}
\caption{Density ratio based adversarial sample detection with varying adversarial sample size}
\begin{algorithmic}
\Require $X$ and $Y$ as datasets with real and adversarial samples with $n$ and $p$ samples in each respectively.
\For{$i \leftarrow 1$ to $t$}
\State $a$: Sample a random index of length 100 without replacement from $n$
\State $b$: Sample a random index of length $m$ without replacement from $p$
\State $x$: Sample from real data $= X[a]$
\State $y$: Sample from adversarial data $= Y[b]$
\State $z$: Sample from real data $= X[b]$
\State Estimate density ratio $R_1 = \frac{p(x)}{p(y)}$
\State Estimate density ratio $R_2 = \frac{p(x)}{p(z)}$
\EndFor
\end{algorithmic}
\end{algorithm}

we use values for $m$ varying from 9 to 1, that is, in a given density ratio estimation, we have 100 real samples with density $p(x)$ and $m$ adversarial samples with density $p(y)$, we also use $m$ real samples with density $p(z)$ for comparison of averaged $R_1$ and $R_2$.

Table 4 shows the results with FGSM created samples with $\epsilon = 0.3$, it is seen that density ratio estimates are capable of detecting adversarial samples even when there is only one adversarial example under investigation.

[24] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In Advances in neural information processing systems, pages 1433–1440, 2008.

[25] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. Neural Networks, 24(7):735–751, 2011.

[26] V. N. Vapnik and V. Vapnik. Statistical learning theory, volume 1. Wiley New York, 1998.
Table 4: Density ratio estimates of real-adversarial and real-real samples using dataset MNIST and FGSM created adversarial samples with varying sample size $m$ and keeping sample size for real images fixed at 100. It is seen that even a single adversarial example can be detected using density ratio estimation with FGSM perturbation of $\epsilon = 0.3$.

| Sample size ($m$) | Real-Adversarial | Real-Real |
|-------------------|-------------------|-----------|
| 9                 | 19.24(18.24,20.25) | 1.55(1.43,1.66) |
| 8                 | 19.84(18.84,20.84) | 1.77(1.57,1.97) |
| 7                 | 20.40(19.29,21.51) | 1.99(1.66,2.33) |
| 6                 | 19.46(18.33,20.59) | 1.79(1.60,1.98) |
| 5                 | 18.90(17.68,20.11) | 1.87(1.63,2.11) |
| 4                 | 17.53(16.22,18.85) | 1.80(1.54,2.05) |
| 3                 | 17.59(16.14,19.03) | 2.15(1.57,2.73) |
| 2                 | 19.90(18.53,21.26) | 2.42(1.93,2.90) |
| 1                 | 22.16(20.92,23.43) | 6.48(5.18,7.77) |