

## 1 A Proof for theorem 3.1

2 Let  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$  denote the train dataset of size  $N$ . Then the minima over the dataset  $\mathcal{D}$ , is  
 3 obtained by solving,

$$\min_{\psi} \mathbb{E}_{\tau \sim U[0,1]} \left[ \frac{1}{N} \sum_i \rho(I[\psi(\mathbf{x}_i, \tau) \geq 0], y_i; \tau) \right] \quad (1)$$

4 Let  $\mathcal{Q}(\mathbf{x}, \tau)$  denotes the solution obtained using the algorithm 1. Let  $\mathcal{P}(\mathbf{x}, \tau)$  denote the solution  
 5 obtained by solving equation 1.

6 We aim to show that  $I[\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0.5] = I[\mathcal{P}(\mathbf{x}_i, \tau) \geq 0]$  for all the points in  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ .

7 First, observe that, since the base classifier  $f_{\theta}(\mathbf{x})$  is obtained using MAE we have that  $I[\mathcal{Q}(\mathbf{x}_i, 0.5) \geq$   
 8  $0.5] = I[f_{\theta}(\mathbf{x}_i) > 0.5] = I[\mathcal{P}(\mathbf{x}_i, 0.5) \geq 0]$ . This is because the loss in equation 1 at  $\tau = 0.5$  is  
 9 nothing but the MAE loss.

10 Next for arbitrary  $\tau$ , we show that  $I[\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0.5] = I[\mathcal{P}(\mathbf{x}_i, \tau) \geq 0]$  over the dataset  $\mathcal{D} =$   
 11  $\{(\mathbf{x}_i, y_i)\}$ .

12 We approximate the indicator function as  $I[\mathbf{x} \geq 0] \approx \lim_{k \rightarrow \infty} K_k(\mathbf{x})$ . For instance one can consider  
 13  $K_k(\mathbf{x}) = \sigma(\mathbf{x}k)$ . Observe that a solution to minimize equation 1 can be obtained by

$$\mathcal{P}(\mathbf{x}, \tau) = \lim_{k \rightarrow \infty} \arg \min_{\psi} \mathbb{E}_{\tau \sim U[0,1]} \left[ \frac{1}{N} \sum_i \rho(K_k(\psi(\mathbf{x}_i, \tau)), y_i; \tau) \right] \quad (2)$$

14 Let

$$\mathcal{P}^{(k)}(\mathbf{x}, \tau) = \arg \min_{\psi} \mathbb{E}_{\tau \sim U[0,1]} \left[ \frac{1}{N} \sum_i \rho(K_k(\psi(\mathbf{x}_i, \tau)), y_i; \tau) \right] \quad (3)$$

15 Also, since  $f(\mathbf{x})$  optimizes MAE, we have for some  $k$ ,  $K_k(\mathcal{P}^{(k)}(\mathbf{x}, 0.5)) = f(\mathbf{x})$ . That is, for some  
 16  $k$ ,

$$\begin{aligned} I[f(\mathbf{x}) \geq 1 - \tau] &= I[K_k(\mathcal{P}^{(k)}(\mathbf{x}, 0.5)) \geq 1 - \tau] \\ &= I[K_k(\mathcal{P}^{(k)}(\mathbf{x}, \tau)) \geq 0.5] \\ &= I[\mathcal{P}^{(k)}(\mathbf{x}, \tau) \geq 0] \end{aligned} \quad (4)$$

17 where the second equality follows from the duality in equation 5. However, for any  $k, k'$ , we have  
 18 that  $I[K_k(\mathcal{P}^{(k)}(\mathbf{x}, \tau)) \geq 0.5]$  is equivalent to  $I[K_{k'}(\mathcal{P}^{(k')}(\mathbf{x}, \tau)) \geq 0.5]$ . Since both  $\mathcal{P}^{(k')}(\mathbf{x}, \tau)$   
 19 and  $\mathcal{P}^{(k)}(\mathbf{x}, \tau)$  would be able to classify the points perfectly at  $\tau$ . So, we have that

$$I[f(\mathbf{x}) \geq 1 - \tau] = I[\mathcal{P}(\mathbf{x}, \tau) \geq 0] \quad (5)$$

20 On the other hand, for all data points in  $\mathcal{D}$  (from the definition of on the construction of  $\mathcal{Q}(\mathbf{x}, \tau)$ ),

$$I[f(\mathbf{x}_i) \geq 1 - \tau] = I[\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0.5] \quad (6)$$

21 Since,  $I[\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0.5] = I[\mathcal{P}(\mathbf{x}_i, \tau) \geq 0]$  for all datapoints in  $\mathcal{D}$ , it follows that  $\mathcal{Q}(\mathbf{x}_i, \tau)$   
 22 optimizes equation 1.

## 23 B Proof for theorem 4.1

24 The proof follows from the fact that

$$\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0 \Leftrightarrow f(\mathbf{x}_i) \geq (1 - \tau) \Leftrightarrow P(g(\mathbf{x}_i) + \epsilon(\mathbf{x}_i) \geq 0) \geq 1 - \tau \quad (7)$$

25 Assuming that  $\tau^* = P(g(\mathbf{x}_i) + \epsilon(\mathbf{x}_i) \geq 0)$ , So, we have

$$\begin{aligned} \int_{\tau=0}^1 I[\mathcal{Q}(\mathbf{x}_i, \tau) \geq 0] d\tau &= \int_{\tau=0}^1 I[\tau^* \geq (1 - \tau)] d\tau = \int_{\tau=0}^1 I[\tau^* \geq (1 - \tau)] d\tau \\ &= \int_{\tau=0}^1 I[\tau \geq (1 - \tau^*)] d\tau = \int_{\tau=(1-\tau^*)}^1 1 d\tau = \tau^* \end{aligned} \quad (8)$$

26 Thus the theorem follows.

Table 1: Comparison of Quantile-Representations with baseline for OOD Detection. Observe that Quantile-Representations outperform the baseline in all the cases.

		DenseNet (Baseline/Quantile-Rep)				
		LSUN(C)	LSUN(R)	iSUN	Imagenet(C)	Imagenet(R)
CIFAR10	AUROC	92.08/93.64	93.86/94.61	92.84/93.74	90.93/91.72	90.93/92.06
	TNR@TPR95	58.19/64.56	63.07/66.89	59.64/64.68	53.94/56.34	54.44/58.22
	Det. Acc	85.58/87.14	87.66/88.60	86.29/87.42	84.11/84.93	84.10/85.33
SVHN	AUROC	91.80/92.29	90.75/90.70	91.21/91.30	91.93/91.97	91.93/92.01
	TNR@TPR95	54.61/58.77	47.67/48.55	48.24/50.15	52.38/53.68	52.43/53.64
	Det. Acc	85.10/85.37	84.32/84.16	84.80/84.77	85.42/85.55	85.46/85.50
		Resnet34 (Baseline/Quantile-Rep)				
		LSUN(C)	LSUN(R)	iSUN	Imagenet(C)	Imagenet(R)
CIFAR10	AUROC	91.43/91.76	92.64/93.08	91.89/92.34	90.59/90.81	89.12/89.39
	TNR@TPR95	54.96/56.76	63.24/65.75	58.56/60.94	52.86/54.89	47.41/49.93
	Det. Acc	84.63/84.96	85.41/86.06	84.39/85.17	83.24/83.44	81.74/82.05
SVHN	AUROC	94.80/94.87	94.37/94.46	95.13/95.22	95.73/95.85	95.62/95.70
	TNR@TPR95	76.19/76.15	72.10/72.87	75.88/76.25	79.16/79.53	78.34/78.82
	Det. Acc	89.58/89.72	88.82/88.87	89.78/89.85	90.72/90.87	90.54/90.60

## 27 C Results when training only the last layer

28 The same observations as done in the main article also hold true when training is done only in the last  
 29 layer by considering the features.

30 **OOD Detection :** These experiments were performed using Densenet and Resnet34 architectures  
 31 on CIFAR10 and SVHN datasets. The OOD datasets are the same as in the main article. Table 1  
 32 shows the results obtained when quantile representations are used only on the last layer.

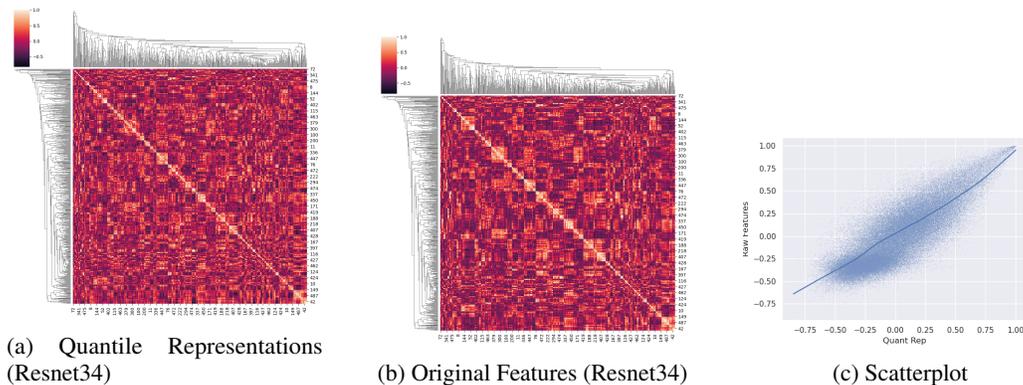


Figure 1: Do quantile representations capture the relevant information for classification? (a) Cross-correlations obtained using Quantile representations for Resnet34 on CIFAR10 (b) Cross-correlations obtained using train features for Resnet34 on CIFAR10. (c) Scatterplot with best fit line (using Locally Weighted Scatterplot Smoothing[1]) of the cross-correlation of features. Observe that as the correlation becomes important (i.e close to  $-1$  or  $1$ ) quantile representations are more consistent with raw features.

33 **Calibration Experiments** The same observations - Quantile probabilities have calibration error  
 34 which does not change with distortion and that these could not be corrected using simple Platt  
 35 Scaling/Isotonic Regression, hold true when training only the last layer as well. This is illustrated in  
 36 figure 3.

37 **Cross-correlation of features** To illustrate that the quantile representations capture the aspects  
 38 of data-distribution relevant to classification, we perform the following experiment - Construct the

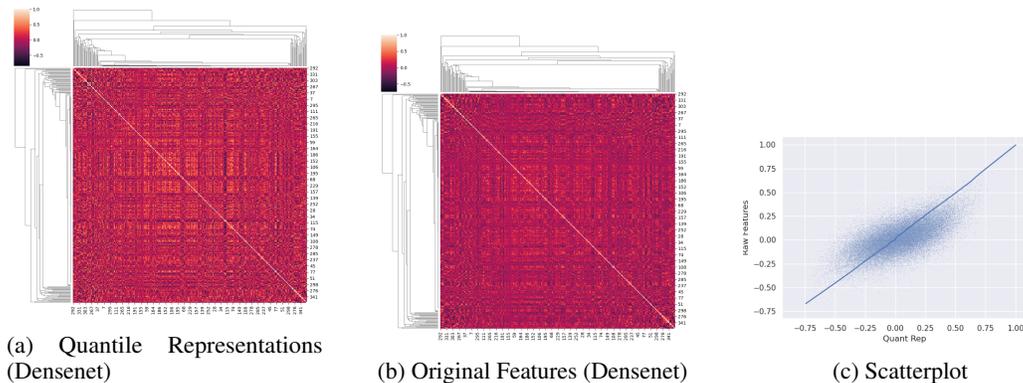


Figure 2: Do quantile representations capture the relevant information for classification? (a) Cross-correlations obtained using Quantile representations for Densenet on CIFAR10 (b) Cross-correlations obtained using train features for Densenet on CIFAR10. (c) Scatterplot with best fit line (using Locally Weighted Scatterplot Smoothing[1]) of the cross-correlations. Observe that as the correlation becomes important (i.e close to  $-1$  or  $1$ ) quantile representations are more consistent with raw features.

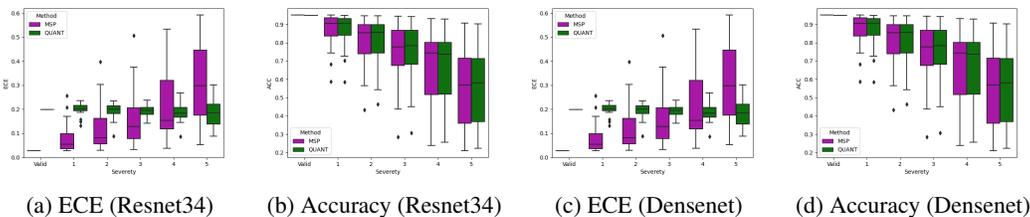


Figure 3: Quantile representations can be effective for calibration because they estimate probabilities using Equation equation 8, which has been shown to be robust to corruptions. As demonstrated using the CIFAR10C dataset [2], the Expected Calibration Error (ECE) of the probabilities obtained from quantile representations (QUANT) does not increase with the severity of the corruptions. In contrast, when using the standard Maximum Softmax Probability (MSP) method, the calibration error increases as the severity of the corruptions increases.

39 cross-correlation between features using (i) Quantile Representations and (ii) Feature values extracted  
 40 using the traindata. If our hypothesis is accurate, then cross-correlations obtained using quantile-  
 41 representations and feature values would be similar.

42 In Figures 1 and 2, we present the results of using features from Resnet34 and Densenet on the  
 43 CIFAR10 dataset. Figures 1a and 1b show the results for Resnet34, and Figures 2a and 2a show the  
 44 results for Densenet. To visualize the cross-correlations, we use a heatmap with row and column  
 45 indices obtained by averaging the linkage of train features. This index is common for both quantile  
 46 representations and extracted features. It is evident from the figure that the cross-correlation between  
 47 features is similar whether it is computed using extracted features or quantile representations.

## 48 D A case where quantile representations do not capture the entire 49 distribution

50 In figure 4 we illustrate an example where quantile representations do not capture the entire distri-  
 51 bution. Here we use the same data as in figure 1, but with different class labels. This is shown in  
 52 figure 4a. When we perform the OOD detection we get the region as in figure 4b. Observe that while  
 53 it does detect points far away from the data as out-of-distribution, the moon structure is not identified.

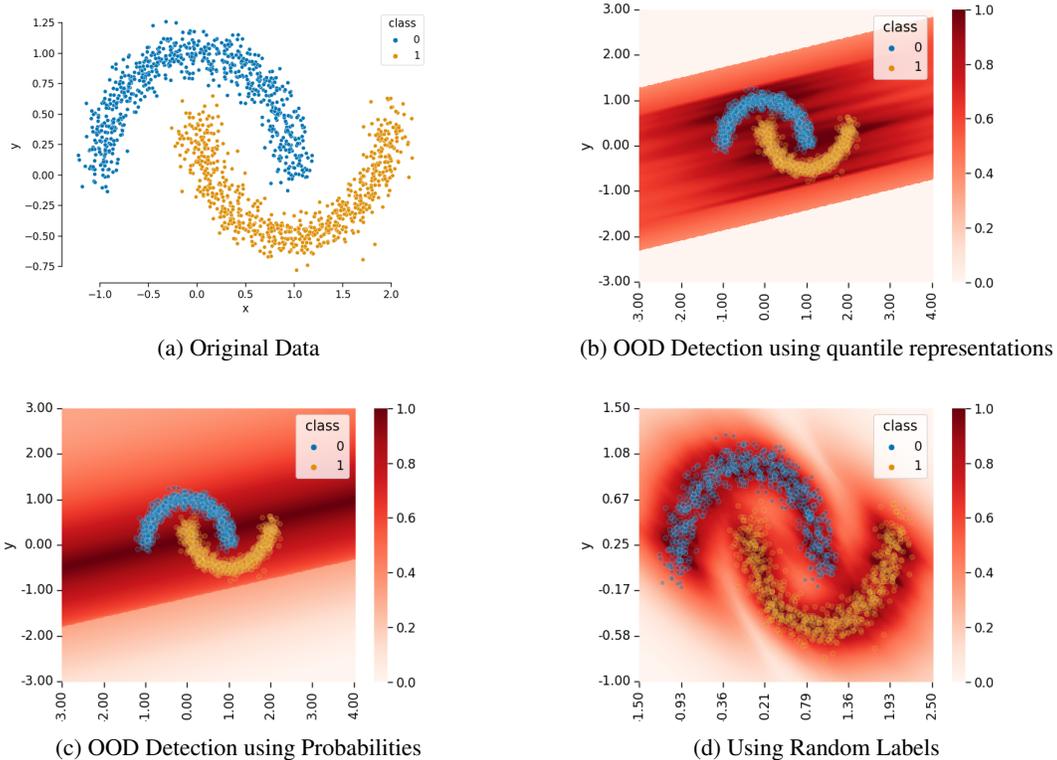


Figure 4: Illustrating a case where quantile representations do not capture the distribution perfectly. (a) Original Dataset. (b) The region detected as in-distribution by using quantile representations. (c) Region detected as in-distribution by using the outputs from a single classifier. Observe that quantile representations still perform better than single classifier outputs. (d) Using random labels instead of ground-truth. Observe that the two moons structure is faithfully preserved in this image. The brightness of **Red** indicates the chance of being in-distribution.

54 In particular, the spaces between the moons is not considered OOD. This illustrates a case when  
 55 quantile representations might fail.

56 However, OOD detection using a single classifier also fail, as illustrated in figure 4c. Observe that  
 57 the region identified by quantile representations is much better than the one obtained using a single  
 58 classifier.

59 **A simple fix for OOD detection:** If OOD detection were the aim, then it is possible to change  
 60 the approach slightly by considering *random labels* instead of the ground-truth labels. This allows  
 61 us to identify arbitrary regions where the data is located. This is illustrated in figure 4d. Observe  
 62 that this method can be used to identify any region in the space by suitably sampling and assigning  
 63 pseudo-labels. In this case, we identify the training data perfectly.

## 64 E Sanity Check - Preserving Monotonicity Property

65 Note that quantile representations obtained by optimizing the simultaneous loss equation 2, should  
 66 follow the monotonicity property -  $Q(\mathbf{x}, \tau_0) \leq Q(\mathbf{x}, \tau_1) \leftrightarrow \tau_0 \leq \tau_1$ . Since our approach is an  
 67 alternate, the quantile representation learnt using algorithm 1 should satisfy this property as well. We  
 68 verify this as follows - Considering the ResNet34 architecture trained on CIFAR10 dataset, we plot  
 69 the *logits* obtained at different quantiles.

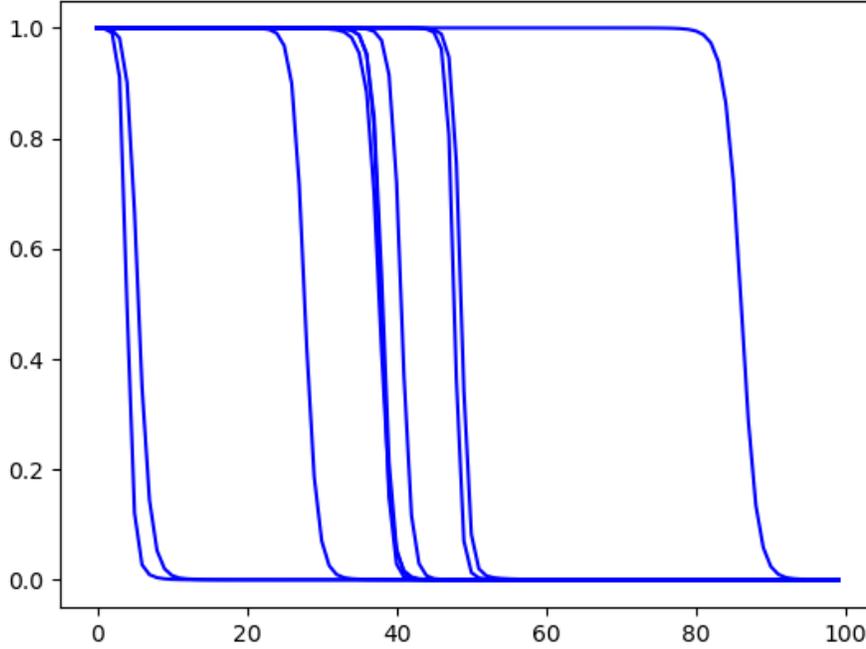


Figure 5: Checking that the quantile representations learnt using algorithm 1 satisfies the monotonicity property.

## 70 F Matching Quantile-Representations to correct the distribution

71 In this part we illustrate the matching of quantile-representations to correct for distribution shifts  
 72 following the ideas from [6]. Let  $X$  denote the original distribution of the data, and let  $\Phi(X)$  denote  
 73 the modified distribution. We assume that the function  $\Phi(\cdot)$  is deterministic but unknown.

74 If both  $X$  and  $\Phi(X)$  are known, then it is easy to estimate  $\Phi(\cdot)$  using some model such as neural  
 75 networks. However, in reality we do not have this information. Once the environment changes, the  
 76 data collected will be very different from the original ones and we do not know how  $\Phi(\cdot)$  distorts the  
 77 original data. So, the aim is to estimate  $\Phi(\cdot)$  without the knowledge of  $X$  and  $\Phi(X)$ . This is where  
 78 the fact that - quantile-representations capture the distribution information becomes relevant.

79 **Is this even possible?** Let  $\mathcal{Q}(x, \tau)$  denote the quantile representation obtained using  $X$ , and  
 80  $\mathcal{Q}_\Phi(x, \tau)$  denote the quantile representation obtained using  $\Phi(X)$ . Let the data collected in the new  
 81 environment be  $\{\hat{x}_i\}$ , then we should have that

$$\int_{\tau=0}^1 |\mathcal{Q}(\Phi^{-}(\hat{x}_i), \tau) - \mathcal{Q}_\Phi(\hat{x}_i, \tau)| = 0 \quad (9)$$

82 Using this it is possible to estimate  $\Phi^{-}$  and hence  $\Phi$ .

83 Observe the following - Functions  $\mathcal{Q}(\cdot, \cdot)$  and  $\mathcal{Q}_\Phi(\cdot, \cdot)$  are learnt from the data using the labels, and  
 84 depends on it. So, one needs the labels to specify the directions in which distribution should be the  
 85 same.

86 For instance, consider the following example - Assume we wish to classify the candidates as  
 87 suitable/not-suitable for a job based on a set of features. Now, what is suitable/not-suitable changes  
 88 with with time. As well as the ability (represented in features) of the general population. So, we  
 89 collect data at time  $t = t_0$ ,  $\{(x_{i,t_0}, y_{i,t_0})\}$  and at time  $t = t_1$ ,  $\{(x_{i,t_1}, y_{i,t_1})\}$ . However we do not

90 know the relation between  $\mathbf{x}_{i,t_0}$  and  $\mathbf{x}_{i,t_1}$ . In such cases, matching quantile representations can be  
 91 useful.

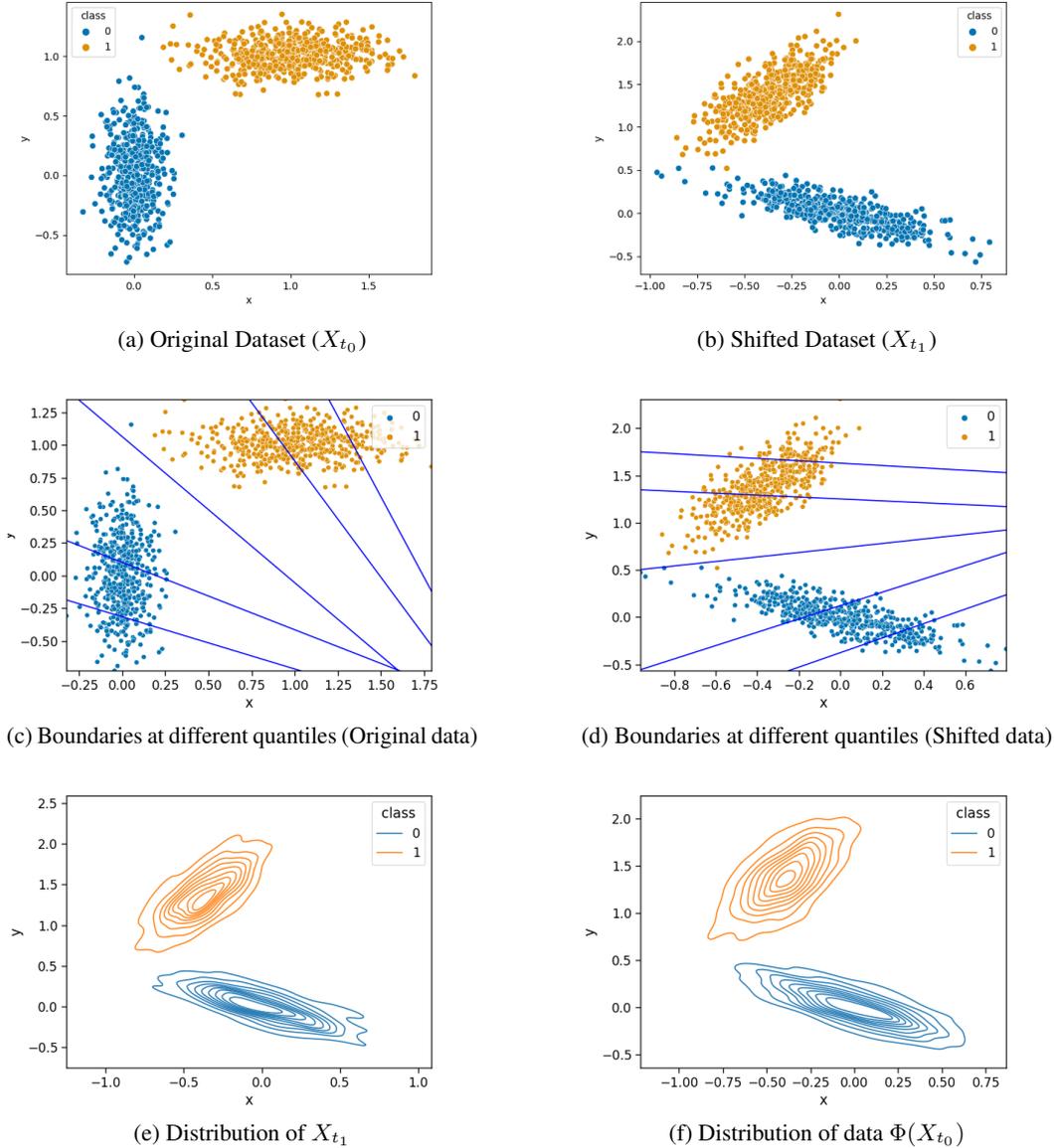


Figure 6: Matching quantile representations. Observe that the estimated distribution at time  $t_1$  is similar to the actual distribution at time  $t_1$ . This shows that the estimate of  $\Phi(\cdot)$  is accurate.

92 **Illustration: Matching of quantile representations** The above procedure is illustrated in figure 6.  
 93 Consider the data at  $t_0$  as in figure 6a and data at  $t_1$  as in figure 6b. This data in figure 6a is generated  
 94 using 2d Gaussian distribution with centers  $[[0, 0], [1, 1]]$  and standard deviation  $[[0.1, 0.3], [0.3, 0.11]]$ .  
 95 We refer to this distribution as  $X_{t_0}$ . Data in figure 6b is obtained by generating a new sample with  
 96 the same distribution as  $X_{t_0}$  and transforming it using a random orthogonal matrix. We refer to this  
 97 distribution using  $X_{t_1}$ . Note that there is no correspondence between the data samples at  $X_{t_0}$  and  
 98  $X_{t_1}$ . Figures 6c and 6d illustrate the quantile representations obtained using the class labels at both  
 99 these times. We then estimate  $\Phi(\cdot)$  using equation 9. Figure 6e shows the density at  $X_{t_1}$  and Figure 6f  
 100 shows the density of  $\Phi(X_{t_0})$ . Observe that the estimate of the density and the actual density match.  
 101 This shows that quantile representations can be used to correct distribution shifts.

102 **Caveat:** However, quantile-representations cannot estimate  $\Phi(\cdot)$  which do not change the distri-  
103 bution of the samples. For instance if  $X_{t_1} = -X_{t_0}$ , and if  $X_{t_0}$  is symmetric around 0, then the  
104 quantile-representations are identical. Under what conditions can  $\Phi(\cdot)$  be estimated is considered for  
105 future work.

106 **Advantage of using quantile-representations** A question which follows is - Why not simply  
107 retrain the classifier at  $t_0$ ? (i) As can be gleaned from the above experiments, it is not possible to  
108 estimate  $\Phi(\cdot)$  from the single classifier alone, but can be done using quantile-representations (ii)  
109 The labels considered for constructing the quantile-representations need not be the same as the  
110 classification labels. They would correspond to important attributes of the data. For instance, one  
111 can consider aspects like technical skill of the candidate instead of simply suitable/not-suitable  
112 classification.

## 113 G Training Details and Compute

114 Training quantile representations was done on a DGX server using 4 GPUs. It takes around 10 hours  
115 (40 GPU hours in total) to learn the quantile representations for each model. We use stochastic  
116 gradient descent with cyclic learning rate for optimization. The base\_lr is taken to be 0.02 and  
117 max\_lr is taken to be 1.0, with exponentially decreasing learning rate using  $\gamma = 0.99994$ . The  
118 batch\_size is taken to be 1024 for resnet34. The number of steps for the cyclic learning is taken to be  
119  $2^{(\text{size\_dataset}/2\text{batch\_size}) + 1}$ . The *size\_dataset* describes the size of the training data.

## 120 H Why Quantile Regression?

121 If the goal of a regression problem is to predict the likely range of estimates (prediction interval) and  
122 not just a single estimate as the Ordinary Least Square Regression (OLS) does, the method is required  
123 to be more general and robust. This method for producing such estimates, relatively unknown in the  
124 Machine Learning community, is known as quantile regression. While OLS regression minimizes  
125 the squared-error loss function to predict a single point estimate, quantile regressions minimize the  
126 quantile loss in predicting a certain quantile. The 50th percentile, otherwise known as the median,  
127 represents the quantile loss as the sum of absolute errors (MAE). Other quantiles could give endpoints  
128 of a prediction interval; for example, a middle-80-percent range is defined by the 10th and 90th  
129 percentiles. The quantile loss differs depending on the evaluated quantile, such that more negative  
130 errors are penalized more for higher quantiles and more positive errors are penalized more for lower  
131 quantiles. In other words, quantile loss varies with the error, depending on the quantile, commonly  
132 interpreted as quantile for under- and over-estimated predictions. The higher the quantile, the more  
133 the quantile loss function penalizes underestimates and the less it penalizes overestimates. Quantiles  
134 allow for an understanding of a probability distribution of a data set in which only the specifications of  
135 the positions are known. Thus, wherever predictions are subject to high uncertainty, quantile should  
136 be the preferred loss function. Quantiles give some information about the shape of a distribution -  
137 in particular whether a distribution is skewed or not; are robust to outliers and can model extreme  
138 events well. Conditional quantiles obtained via regression are used as a robust alternative to classical  
139 conditional means in econometrics and statistics, as they can capture the uncertainty in a prediction,  
140 and model tail behaviors, while making very few distributional assumptions

141 The quantile regression has started relatively recently being applied in the energy-growth nexus  
142 literature. In the past, it has been used extensively in pediatric medicine (offering an optimistic  
143 perspective for precision medicine), survival and duration time studies [3], the determination of  
144 wages, discrimination effects, and income inequality. Also, it has been used in the finance literature  
145 in studies that dealt with bank failure and the time occurrence of this failure [5]. Regarding the more  
146 recent application in the energy-growth nexus field, it is not well documented in the relevant studies  
147 why asymmetries would be present in the way income and wealth is generated in different countries  
148 given the consumption of energy in those countries and other stylized parameters. One reason, quite  
149 understandable, why to use this method, is for testing whether poorer countries will be affected the  
150 same way by energy conservation measures as the rich ones. Another reason as stated by [8] in their  
151 study on renewable energy, oil prices, and economic growth for the United States is that their study  
152 would allow them to determine whether extremely low or high changes in energy consumption prices  
153 would lead economic growth. Therefore we can have very specific and accurate answers to what

154 will happen if there is 1% energy reduction in poor countries. This information would otherwise  
155 have to be included in dummy variables and other forms of robust estimation that assign less weight  
156 to observations that are characterized as outliers. Among the various other statistical twists offered  
157 by the method, the quantile regression may be favored because it does not assume a parametric  
158 distribution and it estimates the entire conditional distribution of the independent variable. Generally,  
159 this method is regarded as more versatile and informative [4].

160 A switch from the squared error to the tilted absolute value loss function allows gradient descent-based  
161 learning algorithms to learn a specified quantile instead of the mean. It means that we can apply  
162 all neural network and deep learning algorithms to quantile regression [3, 5]. The application of  
163 quantiles in deep learning, although relatively recent, are critical for model interpretability. In the  
164 past, [7] extended the notion of conditional quantiles to the binary classification setting—allowing  
165 uncertainty quantification in the predictions, increased resilience to label noise thus furnishing new  
166 insights into the functions learnt by the models. This was accomplished by defining a new loss  
167 called binary quantile regression loss, in the classification setting. The estimated quantiles to obtain  
168 individualized confidence scores provide an accurate measure of a prediction being misclassified.  
169 These scores were then aggregated to compute two additional metrics, namely, confidence score and  
170 retention rate, which can be used to withhold decisions and increase model accuracy. Thus, in a  
171 non-parametric binary quantile classification framework, authors could demonstrate that quantiles  
172 aid in explainability as they can be used to obtain several uni-variate summary statistics that can be  
173 directly applied to existing explanation tools.

174 Therefore, it is not unconvincing to realize the relevance and precedence of quantiles in classification,  
175 in particular, to obtain the conditional quantiles of the underlying latent function learnt by a binary  
176 classifier using customized loss inspired by quantiles [8].

## 177 **References**

- 178 [1] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of*  
179 *the American Statistical Association*, 59(1):829–836, 1979.
- 180 [2] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common  
181 corruptions and perturbations. In *Int. Conf. on Learning Representations*, 2019.
- 182 [3] Q Huang, H Zhang, J Chen, and MJBB He. Quantile regression models and their applications:  
183 A review. *Journal of Biometrics & Biostatistics*, 8(3):1–6, 2017.
- 184 [4] Robert N Rodriguez and Yonggang Yao. Five things you should know about quantile regression.  
185 In *Proceedings of the SAS global forum 2017 conference, Orlando*, pages 2–5, 2017.
- 186 [5] Klaus Schaeck. Bank liability structure, fdic loss, and time to failure: A quantile regression  
187 approach. *Journal of Financial Services Research*, 33:163–179, 2008.
- 188 [6] Nikolaos Sgouropoulos, Qiwei Yao, and Claudia Yastremiz. Matching a distribution by matching  
189 quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759, 2015.  
190 PMID: 26692592.
- 191 [7] Anuj Tambwekar, Anirudh Maiya, Soma S. Dhavala, and Snehanshu Saha. Estimation and  
192 applications of quantiles in deep binary classification. *IEEE Trans. Artif. Intell.*, 3(2):275–286,  
193 2022.
- 194 [8] Victor Troster, Muhammad Shahbaz, and Gazi Salah Uddin. Renewable energy, oil prices, and  
195 economic activity: A granger-causality in quantiles analysis. *Energy Economics*, 70:440–452,  
196 2018.