Appendix A Proof of Lemma 1

Here we show that even in the presence of abstention, learning continues on the true classes. Consider again the loss function defined for a sample x.

$$\mathcal{L}(x) = (1 - p_{k+1}) \left(- \sum_{i=1}^{k} t_i \log \frac{p_i}{1 - p_{k+1}} \right) + \alpha \log \frac{1}{1 - p_{k+1}}.$$

Let j, $(1 \leq j \leq k)$ be the true class for x. During gradient descent, learning on the true class takes place if $\frac{\partial \mathcal{L}}{\partial a_j} < 0$, where a_j is the pre-activation into the softmax unit of class j.

A straight-forward gradient calculation shows that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial a_j} = & -(1 - p_j - p_{k+1}) + p_{k+1} p_j \log \left(\frac{1 - p_{k+1}}{p_j} \right) \\ & - \alpha \frac{p_{k+1} p_j}{1 - p_{k+1}} \end{aligned}$$

Since $\alpha \geq 0$ as per as our assumption, the last quantity in the above expression, $-\alpha \frac{p_{k+1} p_j}{1 - p_{k+1}} \leq 0$

Also note that $(1 - p_j - p_{k+1})$ in the above expression is just the total probability mass in the remaining real (i.e., non-abstention) classes; denote this by q.

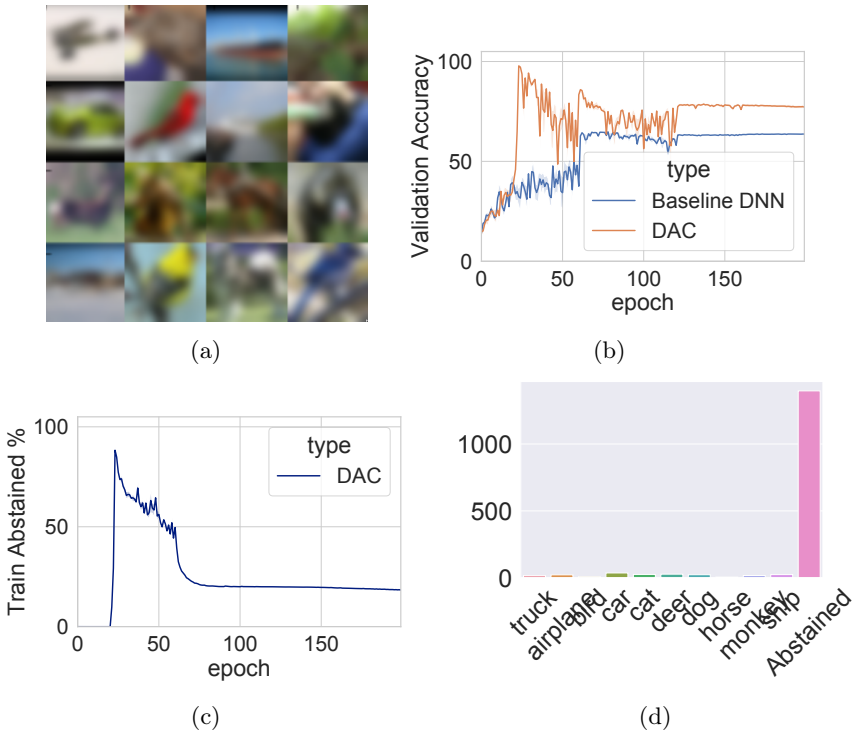
Then we have

$$\begin{aligned} -(1 - p_j - p_{k+1}) + p_{k+1} p_j \log \left(\frac{1 - p_{k+1}}{p_j} \right) & \\ = -q + p_{k+1} p_j \log \left(\frac{1 - p_{k+1}}{p_j} \right) & \\ = -q + p_{k+1} p_j \log \left(\frac{q + p_j}{p_j} \right) & \\ = -q + p_{k+1} p_j \log \left(1 + \frac{q}{p_j} \right) & \\ \leq -q + p_{k+1} p_j \frac{q}{p_j} & \\ = -q + p_{k+1} q & \\ \leq 0 & \end{aligned}$$

where, in A, we have made use of the fact that $\log (1 + x) \leq x$ for all $x > -1$. Thus $\frac{\partial \mathcal{L}}{\partial a_j} \leq 0$ as desired.

Appendix B Noisy Labels associated with a data transformation

We present further results on the abstaining ability of the DAC in the presence of structured noise. Here we simulate a scenario where a subset of the training



(a) (b) (c) (d)

Figure 1: Results on blurred-image experiment with noisy labels (a)20% of the images are blurred in the train set, and their labels randomized (b) Validation accuracy for baseline vs DAC (non-abstained) (c)Abstention behavior for the DAC during training (d) Distribution of predictions on the blurred validation images for the DAC. We also observed (not shown) that for the baseline DNN, the accuracy on the blurred images in the validation set is no better than random.

data, due to feature degradation, ends up with unreliable labels. We apply a Gaussian blurring transformation to 20% of the train and test images across all the classes (Figure 1a), and randomize the labels on the blurred training set. This is similar to the smudging experiment, but lacks the presence of a consistent, conspicuous feature that the DAC can associate with abstention. On the other hand, the lack of high frequency components, or conversely the abundance of low frequency components, might itself be thought of as a feature that is consistent across the samples that have had their label randomized. **Results** The DAC abstains remarkably well on the blurred images in the test set (Figure 1d), while maintaining classification accuracy over the remaining samples in the validation set ($\approx$ 79%). The baseline DNN accuracy drops to 63% (Figure 1b), while the baseline accuracy over the smudged images alone is no better than random ($\approx$ 9.8%) . The abstention behavior of the DAC on the blurred images in the test set can be explained by how abstention evolves during training (Figure 1c). Once

abstention is introduced at epoch 20, the DAC initially opts to abstain on a high percentage of the training data, while continuing to learn (since the gradients w.r.t the true-class pre-activations are always negative.). In the later epochs, sufficient learning has taken place on the non-randomized samples but the DAC continues to abstain on about 20% of the training data, which corresponds to the blurred images indicating that a strong association has been made between blurring and abstention.

Appendix C Results on Non-Uniform Label Noise

Dataset	Method	Class Dependent Label Noise Fraction			
		η =0.1	0.2	0.3	0.4
CIFAR-10 (ResNet-34)	$\mathcal{L}_q$	90.91	89.33	85.45	76.74
	Trunc $\mathcal{L}_q$	90.43	89.45	87.10	82.28
	Forward T	91.32	90.35	89.25	88.12
	Forward $\hat{T}$	90.52	89.09	86.79	83.55
	DAC	**94.23**	**93.20**	**92.07**	**89.88**
CIFAR-100 (ResNet-34)	$\mathcal{L}_q$	68.36	66.59	61.45	47.22
	Trunc $\mathcal{L}_q$	68.86	66.59	61.87	47.66
	Forward T	71.05	71.08	70.76	**70.82**
	Forward $\hat{T}$	45.96	42.46	38.13	34.44
	DAC	**75.59**	**73.22**	**71.38**	65.34
Fashion-MNIST (ResNet-18)	$\mathcal{L}_q$	93.51	93.24	92.21	89.53
	Trunc $\mathcal{L}_q$	93.53	93.36	92.76	91.62
	Forward T	94.33	94.03	93.91	93.65
	Forward $\hat{T}$	94.09	93.66	93.52	88.53
	DAC	**95.48**	**95.08**	**94.96**	**94.31**

Table 1: Comparison of DAC vs related methods for class-dependent label noise. Performance numbers reproduced from (Zhang & Sabuncu, 2018). For the DAC, an abstaining classifier is first used to identify and eliminate label noise, and an identical DNN is then used for downstream training.

Here we report results on CIFAR-10, CIFAR-100 and Fashion-MNIST for class-dependent label noise. The experimental setup is exactly as described in (Zhang & Sabuncu, 2018), and we compare with the results reported in that paper which also includes the Forward correction method of (Patrini et al., 2017). In CIFAR-10, the class dependent noise results in the following flip scenario with probability η: TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, and CAT $\leftrightarrow$ DOG with probability. For CIFAR-100, classes are organized into groups as described in (Krizhevsky & Hinton, 2009) and the class-dependent noise is simulated by flipping each class into the next circularly with probability η. For Fashion-MNIST, classes are flipped as follows: BOOT

$\rightarrow$ SNEAKER , SNEAKER $\rightarrow$ SANDALS, PULLOVER $\rightarrow$ SHIRT, COAT$\rightarrow$ DRESS with probability η. The aforementioned flipping scenarios are identical to the setup in (Zhang & Sabuncu, 2018). Results are shown in Table 1.

References

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pp. 2233–2241, 2017.

Zhang, Z. and Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*, 2018.