Does Label Smoothing Mitigate Label Noise?

Michal Lukasik [1] **Srinadh Bhojanapalli** [1] **Aditya Krishna Menon** [1] **Sanjiv Kumar** [1]

Abstract

Label smoothing is commonly used in training deep learning models, wherein one-hot training labels are mixed with uniform label vectors. Empirically, smoothing has been shown to improve both predictive performance and model calibration. In this paper, we study whether label smoothing is also effective as a means of coping with *label noise*. While label smoothing apparently *amplifies* this problem — being equivalent to injecting symmetric noise to the labels — we show how it relates to a general family of *loss-correction* techniques from the label noise literature. Building on this connection, we show that label smoothing is competitive with loss-correction under label noise. Further, we show that when distilling models from noisy data, label smoothing of the teacher is beneficial; this is in contrast to recent findings for noise-free problems, and sheds further light on settings where label smoothing is beneficial.

1. Introduction

Label smoothing is commonly used to improve the performance of deep learning models (Szegedy et al., 2016; Chorowski & Jaitly, 2017; Vaswani et al., 2017; Zoph et al., 2018; Real et al., 2018; Huang et al., 2019; Li et al., 2020). Rather than standard training with one-hot training labels, label smoothing prescribes using *smoothed* labels by mixing in a uniform label vector. This procedure is generally understood as a means of regularisation (Szegedy et al., 2016; Zhang et al., 2018) that improves generalization and model calibration (Pereyra et al., 2017; Müller et al., 2019).

How does label smoothing affect the robustness of deep networks? Such robustness is desirable when learning from data subject to *label noise* (Angluin & Laird, 1988). Modern deep networks can perfectly fit such noisy labels (Zhang et al., 2017). Can label smoothing address this problem? Interestingly, there are two competing intuitions. On the one hand, smoothing might *mitigate* the problem, as it prevents overconfidence on any one example. On the other hand, smoothing might *accentuate* the problem, as it is equivalent to injecting uniform noise into all labels (Xie et al., 2016).

Which of these intuitions is borne out in practice? A systematic study of this question is, to our knowledge, lacking. Indeed, label smoothing is conspicuously absent in most treatments of the noisy label problem (Patrini et al., 2016; Han et al., 2018b; Charoenphakdee et al., 2019; Thulasidasan et al., 2019; Amid et al., 2019; Menon et al., 2020). Intriguingly, however, a cursory inspection at popular *loss correction* techniques in this literature (Natarajan et al., 2013; Patrini et al., 2017; van Rooyen & Williamson, 2018) reveals a strong similarity to label smoothing (see §3). But what is the precise relationship between these methods, and does it imply label smoothing is a viable denoising technique?

In this paper, we address these questions by first connecting label smoothing to existing label noise techniques. At first glance, this connection indicates that smoothing has an *opposite* effect to one such *loss-correction* technique. However, we empirically show that smoothing is competitive with such techniques in denoising, and that it improves performance of *distillation* (Hinton et al., 2015) under label noise. We then explain its denoising ability by analysing smoothing as a *regulariser*. In sum, our contributions are:

(i) we present a novel connection of label smoothing to loss correction techniques from the label noise literature (Natarajan et al., 2013; Patrini et al., 2017).

(ii) we empirically demonstrate that label smoothing significantly improves performance under label noise at varying noise levels, and is competitive with loss correction techniques. Our experiments show that smoothing improves accuracy on both the clean and noisy parts of the data, while preserving model calibration. We explain these denoising effects by relating label smoothing to ℓ_2 *regularisation*.

(iii) we show that when distilling from noisy labels, smoothing the teacher *improves* the student; this is in marked contrast to recent findings in noise-free settings.

Contributions (i) and (ii) establish that label smoothing can be beneficial under noise, and also highlight that a *regularisation* view can complement a *loss* view, the latter

[1]Google Research. Correspondence to: Michal Lukasik <mlukasik@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

being more popular in the noise literature (Patrini et al., 2017). Contribution (iii) continues a line of exploration initiated in Müller et al. (2019) as to the relationship between teacher accuracy and student performance. While Müller et al. (2019) established that label smoothing can *harm* distillation, we show an *opposite* picture in noisy settings.

2. Background and Notation

We present some background on (noisy) multiclass classification, label smoothing, and knowledge distillation.

2.1. Multiclass Classification

In multiclass classification, we seek to classify instances $\mathcal{X}$ into one of L labels $\mathcal{Y} = [L] \doteq \{1, 2, \dots, L\}$. More precisely, suppose instances and labels are drawn from a distribution $\mathbb{P}$. Let $\ell \colon [L] \times \mathbb{R}^L \to \mathbb{R}_+$ be a *loss* function, where $\ell(y, \mathbf{f})$ is the penalty for predicting *scores* $\mathbf{f} \in \mathbb{R}^L$ given true label $y \in [L]$. We seek a predictor $\mathbf{f} \colon \mathcal{X} \to \mathbb{R}^L$ minimising the *risk* of $\mathbf{f}$, i.e., its expected loss under $\mathbb{P}$:

$$R(\mathbf{f}) \doteq \mathop{\mathbb{E}}_{(x,y)} \left[\ell(y, \mathbf{f}(x))\right] = \mathop{\mathbb{E}}_{x} \left[\mathbf{p}^*(x)^{\mathrm{T}} \ell(\mathbf{f}(x))\right],$$

where $\mathbf{p}^*(x) \doteq \left[\mathbb{P}(y \mid x)\right]_{y \in [L]}$ is the class-probability distribution, and $\ell(\mathbf{f}) \doteq \left[\ell(y, \mathbf{f})\right]_{y \in [L]}$. Canonically, ℓ is the softmax cross-entropy, $\ell(y, \mathbf{f}) \doteq -f_y + \log \sum_{y' \in [L]} e^{f_{y'}}$.

Given a finite training sample $S = \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$, one can minimise the *empirical risk*

$$R(\mathbf{f}; S) \doteq \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, \mathbf{f}(x_n)).$$

In *label smoothing* (Szegedy et al., 2016), one mixes the training labels with a uniform mixture over all possible labels: for $\alpha \in [0, 1]$, this corresponds to minimising

$$\bar{R}(\mathbf{f}; S) = \frac{1}{N} \sum_{n=1}^{N} \bar{\mathbf{y}}_n^{\mathrm{T}} \ell(\mathbf{f}(x_n)), \tag{1}$$

where $(\bar{\mathbf{y}}_n)_i \doteq (1 - \alpha) \cdot [\![i = y]\!] + \frac{\alpha}{L}$.

2.2. Learning under Label Noise

The *label noise* problem is the setting where one observes samples from some distribution $\bar{\mathbb{P}}$ with $\bar{\mathbb{P}}(y \mid x) \neq \mathbb{P}(y \mid x)$; i.e., the observed labels are not reflective of the ground truth (Angluin & Laird, 1988; Scott et al., 2013). Our goal is to nonetheless minimise the risk on the (unobserved) $\mathbb{P}$. This poses a challenge to deep neural networks, which can fit completely arbitrary labels (Zhang et al., 2017).

A common means of coping with noise is to posit a noise model, and design robust procedures under this model. One simple model is *class-conditional noise* (Blum & Mitchell, 1998; Scott et al., 2013; Natarajan et al., 2013), wherein there is a row-stochastic *noise transition* matrix $\mathbf{T} \in [0, 1]^{L \times L}$ such that for each $(x, y) \sim \mathbb{P}$, label y may be flipped to y' with probability $\mathbf{T}_{y,y'}$. Formally, if $\bar{\mathbf{p}}_y^*(x) \doteq \bar{\mathbb{P}}(y \mid x)$ and $\mathbf{p}_y^*(x) \doteq \mathbb{P}(y \mid x)$ are the noisy and clean class-probabilities respectively, we have

$$\bar{\mathbf{p}}^*(x) = \mathbf{T}^{\mathrm{T}} \mathbf{p}^*(x). \tag{2}$$

The *symmetric noise* model further assumes that there is a constant flip probability $\rho \in \left[0, 1 - \frac{1}{L}\right)$ of changing the label uniformly to one of the other classes (Long & Servedio, 2010; van Rooyen et al., 2015), i.e., for $\alpha \doteq \frac{L}{L-1} \cdot \rho$,

$$\mathbf{T} = (1 - \alpha) \cdot \mathbf{I} + \frac{\alpha}{L} \cdot \mathbf{J} \tag{3}$$

where $\mathbf{I}$ denotes the identity and $\mathbf{J}$ the all-ones matrix.

While there are several approaches to coping with noise, our interest will be in the family of *loss correction* techniques: assuming one has knowledge (or estimates) of the noise-transition matrix $\mathbf{T}$, such techniques yield consistent risk minimisers with respect to $\mathbb{P}$. (Patrini et al., 2017) proposed two such techniques, termed *backward* and *forward correction*, which respectively involve the losses

$$\ell^{\leftarrow}(\mathbf{f}) = \mathbf{T}^{-1} \ell(\mathbf{f}) \tag{4}$$

$$\ell^{\rightarrow}(\mathbf{f}) = \ell(\mathbf{T}\mathbf{f}). \tag{5}$$

Observe that for a given label y, $\ell^{\leftarrow}(y, \mathbf{f}) = \sum_{y' \in [L]} T^{-1}_{yy'} \cdot \ell(y', \mathbf{f}(x))$ computes a weighted sum of *losses* for all labels $y' \in [L]$, while $\ell^{\rightarrow}(y, \mathbf{f}) = \ell\left(y, \sum_{y' \in [L]} T_{:y'} \cdot f_{y'}(x)\right)$ computes a weighted sum of *predictions* for all $y' \in [L]$.

Backward correction was inspired by techniques in Natarajan et al. (2013); Cid-Sueiro et al. (2014); van Rooyen & Williamson (2018), and results in an unbiased estimate of the risk with respect to $\mathbb{P}$. Recent works have studied robust estimation of the $\mathbf{T}$ matrix from noisy data alone (Patrini et al., 2017; Han et al., 2018b; Xia et al., 2019). Forward correction was inspired by techniques in Reed et al. (2014); Sukhbaatar et al. (2015), and does *not* result in an unbiased risk estimate. However, it preserves the Bayes-optimal minimiser, and is empirically effective (Patrini et al., 2017).

2.3. Knowledge Distillation

Knowledge distillation (Bucilă et al., 2006; Hinton et al., 2015) refers to the following recipe: given a training sample $S \sim \mathbb{P}^N$, one trains a *teacher* model using a loss function suitable for estimating class-probabilities, e.g., the softmax cross-entropy. This produces a class-probability estimator $\mathbf{p}^{\mathrm{t}} \colon \mathcal{X} \to \Delta_L$, where Δ denotes the simplex. One then uses $\{(x_n, \mathbf{p}^{\mathrm{t}}(x_n))\}_{n=1}^N$ to train a *student* model, e.g., using cross entropy (Hinton et al., 2015) or square loss (Sanh

et al., 2019) as an objective. The key advantage of distillation is that the resulting student has improved performance compared to simply training the student on labels in S.

3. Label Smoothing Meets Loss Correction

We now relate label smoothing to loss correction techniques for label noise via a *label smearing* framework.

3.1. Label Smearing for Loss Functions

Suppose we have some base loss ℓ of interest, e.g., the softmax cross-entropy. Recall that we summarise the loss via the vector $\ell(\mathbf{f}) \doteq \left[\ell(y, \mathbf{f})\right]_{y \in [L]}$. The loss on an example (x, y) is $\ell(y, \mathbf{f}(x)) = \mathbf{e}_y^{\mathrm{T}} \ell(\mathbf{f}(x))$ for one-hot vector $\mathbf{e}_y$.

Consider now the following generalisation, which we term *label smearing*: given a matrix $\mathbf{M} \in \mathbb{R}^{L \times L}$, we compute

$$\ell^{\mathrm{SM}}(\mathbf{f}) \doteq \mathbf{M}\, \ell(\mathbf{f}).$$

On an example (x, y), the *smeared loss* is given by

$$\mathbf{e}_y^{\mathrm{T}} \ell^{\mathrm{SM}}(\mathbf{f}(x)) = M_{yy} \cdot \ell(y, \mathbf{f}(x)) + \sum_{y' \neq y} M_{yy'} \cdot \ell(y', \mathbf{f}(x)).$$

Compared to the standard loss, we now potentially involve *all* possible labels, scaled appropriately by the matrix $\mathbf{M}$.

3.2. Special Cases of Label Smearing

The label smearing framework captures many interesting approaches as special cases (see Table 1):

- *Standard training*. Suppose that $\mathbf{M} = \mathbf{I}$, for identity matrix $\mathbf{I}$. This trivially corresponds to standard training.
- *Label smoothing*. Suppose that $\mathbf{M} = (1 - \alpha) \cdot \mathbf{I} + \frac{\alpha}{L} \cdot \mathbf{J}$, where $\mathbf{J}$ is the all-ones matrix, and $\alpha \in [0, 1]$ is a tuning parameter. This corresponds to mixing the true label with a uniform distribution over all the classes, which is precisely label smoothing per (1).
- *Backward correction*. Suppose that $\mathbf{M} = \mathbf{T}^{-1}$, where $\mathbf{T}$ is a class-conditional noise transition matrix. This corresponds to the backward correction procedure of Patrini et al. (2017). Here, the entries of $\mathbf{M}$ may be *negative*; indeed, for symmetric noise, $\mathbf{M} = \frac{1}{1-\alpha} \cdot \left(\mathbf{I} - \frac{\alpha}{L} \cdot \mathbf{J}\right)$ where $\alpha \doteq \frac{L}{L-1} \cdot \rho$. While this poses optimisation problems, recent works have studied means of correcting this (Kiryo et al., 2017; Han et al., 2018a).

The above techniques have been developed with different motivations. By casting them in a common framework, we can elucidate some of their shared properties.

3.3. Statistical Consistency of Label Smearing

Recall that our fundamental goal is to devise a procedure that can approximately minimise the population risk $R(\mathbf{f})$.

Method	Smearing matrix
Standard	$\mathbf{I}$
Label smoothing	$(1 - \alpha) \cdot \mathbf{I} + \frac{\alpha}{L} \cdot \mathbf{J}$
Backward correction	$\frac{1}{1-\alpha} \cdot \mathbf{I} - \frac{\alpha}{(1-\alpha) \cdot L} \cdot \mathbf{J}$

Table 1. Comparison of different label smearing methods. Here, $\mathbf{I}$ denotes the identity and $\mathbf{J}$ the all-ones matrix. For backward correction, the theoretical optimal choice of $\alpha = \frac{L}{L-1} \cdot \rho$, where ρ is the level of symmetric label noise.

Given this, it behooves us to understand the effect of label smearing on this risk. As we shall explicate, label smearing:

(i) is equivalent to fitting to a modified distribution.

(ii) preserves classification consistency for suitable $\mathbf{M}$.

For (i), observe that the smeared loss has corresponding risk

$$\begin{aligned} R_{\mathrm{sm}}(\mathbf{f}) &= \mathop{\mathbb{E}}_{x} \left[\mathbf{p}^*(x)^{\mathrm{T}} \ell^{\mathrm{SM}}(\mathbf{f}(x))\right] \\ &= \mathop{\mathbb{E}}_{x} \left[\mathbf{p}^*(x)^{\mathrm{T}} \mathbf{M}\, \ell(\mathbf{f}(x))\right]. \end{aligned}$$

Consequently, minimising a smeared loss is equivalent to minimising the original loss on a *smeared* distribution with class-probabilities $\mathbf{p}^{\mathrm{SM}}(x) = \mathbf{M}^{\mathrm{T}} \mathbf{p}^*(x)$.

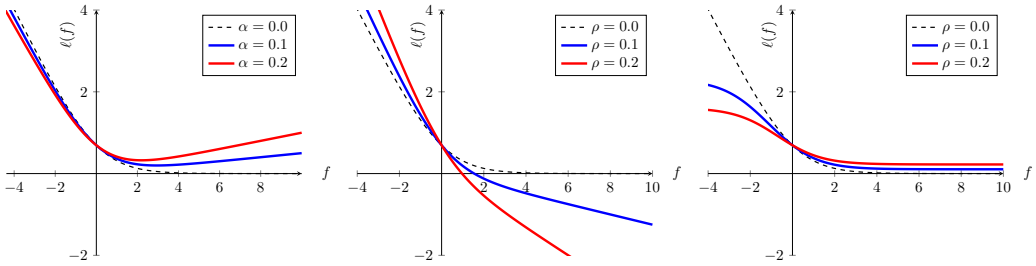
For example, under label smoothing, we fit to the class-probabilities $\mathbf{M}^{\mathrm{T}} \mathbf{p}^*(x) = (1 - \alpha) \cdot \mathbf{p}^*(x) + \frac{\alpha}{L}$. This corresponds to a scaling and translation of the original. This trivially preserves the label with maximal probability, provided $\alpha < 1$. Smoothing is thus *consistent* for classification, i.e., minimising its risk also minimises the classification risk (Zhang, 2004a;b; Bartlett et al., 2006).

Now consider backward correction with $\mathbf{M} = \mathbf{T}^{-1}$. Suppose this is applied to a distribution with class-conditional label noise governed by transition matrix $\mathbf{T}$. Then, we will fit to probabilities $\mathbf{M}^{\mathrm{T}} \bar{\mathbf{p}}^*(x) = (\mathbf{T}^{\mathrm{T}})^{-1} \bar{\mathbf{p}}^*(x)$. By (2), these will exactly equal the *clean* probabilities $\mathbf{p}^*(x)$; i.e., backward correction will effectively denoise the labels.

3.4. Relating Label Smoothing and Loss Correction

Following Table 1, one cannot help but notice a strong similarity between label smoothing and backward correction for symmetric noise. Both methods combine an identity matrix with an all-ones matrix; the striking difference, however, is that this combination is via *addition* in one, but *subtraction* in the other. This results in losses with very different forms:

$$\begin{aligned} \ell^{\mathrm{LS}}(y, \mathbf{f}) &\propto \ell(y, \mathbf{f}) + \frac{\alpha}{(1 - \alpha) \cdot L} \cdot \sum_{y'} \ell(y', \mathbf{f}) \qquad (6) \\ \ell^{\leftarrow}(y, \mathbf{f}) &\propto \ell(y, \mathbf{f}) - \frac{\alpha}{L} \sum_{y'} \ell(y', \mathbf{f}). \end{aligned}$$



(a) Label smoothing. (b) Backward correction. (c) Forward correction.

Figure 1. Effect of label smoothing, backward correction, and forward correction on the logistic loss. The standard logistic loss vanishes for large positive predictions, and is linear for large negative predictions. Smoothing introduces a finite positive minima. Backward correction makes the loss *negative* for large positive predictions. Forward correction makes the loss *saturate* for large negative predictions.

Fundamentally, the effect of the two techniques is different: smoothing aims to *minimise* the average per-class loss $\frac{1}{L}\sum_{y'} \ell(y', \mathbf{f})$, while backward correction seeks to *maximise* this. Figure 1 visualises the effect on the losses when $L = 2$, and ℓ is the logistic loss. Intriguingly, the smoothed loss is seen to penalise confident predictions. On the other hand, backward correction allows one to compensate for overly confident negative predictions by allowing for a *negative* loss on positive samples that are correctly predicted.

Label smoothing also relates to forward correction: recall that here, we compute the loss $\ell^{\rightarrow}(\mathbf{f}) = \ell(\mathbf{T}\mathbf{f})$. Compared to label smoothing, forward correction thus performs smoothing of the *logits*. As shown in Figure 1, the effect is that the loss becomes bounded for all predictions.

At this stage, we return to our original motivating question: can label smoothing mitigate label noise? The above would seem to indicate otherwise: backward correction guarantees an unbiased risk estimate, and yet we have seen smoothing constructs a fundamentally different loss. In the next section, we assess whether this is borne out empirically.

4. Effect of Label Smoothing on Label Noise

We now present experimental observations of the effects of label smoothing under label noise. We then provide insights into why smoothing can successfully denoise labels, by viewing smoothing as a form of *shrinkage regularisation*.

4.1. Denoising Effects of Label Smoothing

We begin by empirically answering the question: can label smoothing successfully mitigate label noise? To study this, we employ smoothing in settings where the training data is artificially injected with symmetric label noise. This follows the convention in the label noise literature (Patrini et al., 2017; Han et al., 2018a; Charoenphakdee et al., 2019).

Specifically, we consider the CIFAR-10, CIFAR-100 and ImageNet datasets, and add symmetric label noise at level $\rho^* = 20\%$ to the training (but *not* the test) set; i.e., we replace the training label with a uniformly chosen label 20% of the time. On CIFAR-10 and CIFAR-100 we train two different models on this noisy data, ResNet-32 and ResNet-56, with similar hyperparameters as Müller et al. (2019). Each experiment is repeated five times, and we report the mean and standard deviation of the *clean* test accuracy. On ImageNet we train ResNet-v2-50 with LARS (You et al., 2017). We describe the detailed experimental setup in Appendix B.

As loss functions, our baseline is training with the softmax cross-entropy on the noisy labels. We then employ label smoothing (1) (**LS**) for various values of α, as well as forward (**FC**) and backward (**BC**) correction (4), (5) assuming symmetric noise for various values of α. We remark here that in the label noise literature, it is customary to *estimate* α, with theoretical optimal value $\alpha^* = \frac{L}{L-1} \cdot \rho^*$; however, we shall here simply treat this as a tuning parameter akin to the smoothing α, whose effect we shall study.

We now analyse the results along several dimensions.

Accuracy: In Figure 2, we plot the test accuracies of all methods on CIFAR-10 and CIFAR-100. Our first finding is that *label smoothing significantly improves accuracy over the baseline*. We observe similar denoising effects on ImageNet in Table 2. Moreover, a similar trend holds for higher levels of label noise, as reported in Table 3. This confirms that empirically, label smoothing is effective in dealing with label noise.

Our second finding is that, surprisingly, *choosing* $\alpha \gg \rho^*$, *the true noise rate, improves performance of all methods*. This is in contrast to the theoretically optimal choice $\alpha \approx \rho^*$ for loss correction approaches (Patrini et al., 2017), and indicates it is valuable to treat α as a tuning parameter.

Finally, we see that label smoothing is *often competitive with loss correction*. This is despite it minimising a fundamentally different loss to the unbiased backward correction,

	0.0	0.1	0.2	0.4	0.6
LS	70.86	71.12	71.55	70.95	70.59
FC	70.86	73.04	73.17	73.35	72.92

Table 2. Test accuracy on ImageNet trained with $\rho = 20\%$ label noise on ResNet-v2-50, with label smoothing (LS) and forward correction (FC) for varying α. Both LS and FC successfully denoise, and thus improve over the baseline ($\alpha = 0$).

ρ	0.0	0.2	0.4	0.6	0.8
40%	50.67	52.86	55.90	56.49	54.98
60%	37.58	39.43	42.39	44.38	42.35

Table 3. Test accuracy on CIFAR-100 trained with $\rho = 40\%$ and $\rho = 60\%$ label noise with label smoothing (LS) for varying α. LS improves over the baseline ($\alpha = 0$). Results averaged over 3 runs.

as discussed in §3.4. We note however that loss correction generally produces the best overall accuracy with high α.

Denoising: What explains the effectiveness of label smoothing for training with label noise? Does it correct the predictions on noisy examples, or does it only further improve the predictions on the clean (non-noisy) examples?

To answer these questions, we separately inspect accuracies on the noisy and clean portions of the training data (i.e., on those samples whose labels are flipped, or not). Table 4 reports this breakdown from the ResNet-32 model on CIFAR-100, for different values of α. We see that as α increases, accuracy improves on both the noisy and clean parts of the data, with a more significant boost on the noisy part. Consequently, smoothing *systematically improves predictions* of both clean and noisy samples.

Model Confidence: Predictive accuracy is only concerned with a model ranking the true label ahead of the others. However, the *confidence* in model predictions is also of interest, particularly since a danger with label noise is being overly confident in predicting a noisy label. How do smoothing and correction methods affect this confidence under noise?

To measure this, in Figure 3 we plot distributions of the differences between the logit activation $\hat{p}(y \mid x)$ for a true/noisy label y, and the average logit activation $\frac{1}{L}\sum_{y' \in [L]} \hat{p}(y' \mid x)$. Compared to the baseline, label smoothing significantly *reduces confidence* in the *noisy* label (refer to the left side of Figure 3(b)).

To visualise this effect of smoothing, in Figure 4 we plot pre-logits (penultimate layer output) of examples from 3 classes projected onto their class vectors as in Müller et al. (2019), for a ResNet-32 trained on CIFAR-100. As we increase α, the confidences for noisy labels shrink, showing the denoising effects of label smoothing.

α	**Full train**	**Clean part**	**Noisy part**	
	true labels	true labels	true labels	noisy labels
0.0	77.39	86.75	39.92	17.88
0.1	80.11	87.99	48.58	12.27
0.2	81.22	88.27	53.01	8.32

Table 4. Accuracy on different portions of the training set from ResNet-32, trained with different label smoothing values α on CIFAR-100. As α increases, accuracy improves on both clean and noisy part of data. Interestingly, the improvement on the noisy part of data is greater than the reduction in fit to the noisy labels (compare the two rightmost columns in the table). Thus, there are noisy examples assigned neither to correct class nor to the observed *noisy* class without LS, and which LS helps classify correctly.

α	**LS**	**FC**	**BC**
0.0	0.111	0.111	0.111
0.1	0.108	0.153	0.214
0.2	0.156	0.165	0.266

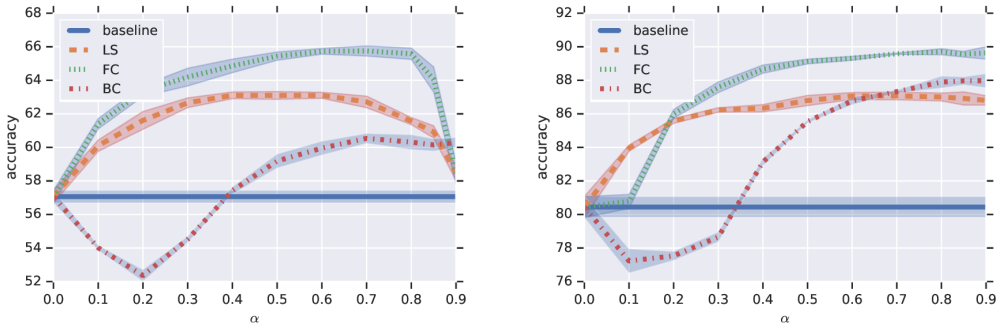
Table 5. Expected calibration error (ECE) computed on 100 bins on test set for ResNet-32 on CIFAR-100, trained with different label smearing techniques under varying values of α. Generally, label smearing is detrimental to calibration.

On the other hand, both forward and backward correction systematically *increase confidence* in predictions. This is especially pronounced for forward correction, demonstrated by the large spike for high differences in Figure 3(b). At the same time, these techniques increase the confidence in predictions of the true label (refer to Figure 3(a)): forward correction in particular becomes much more confident in the true label than any other technique.

In sum, Figure 3 illustrates both positive and adverse effects on confidence from label smearing techniques: label smoothing becomes less confident in both the noisy and correct labels, while forward and backward correction become more confident in both the correct labels and noisy labels.

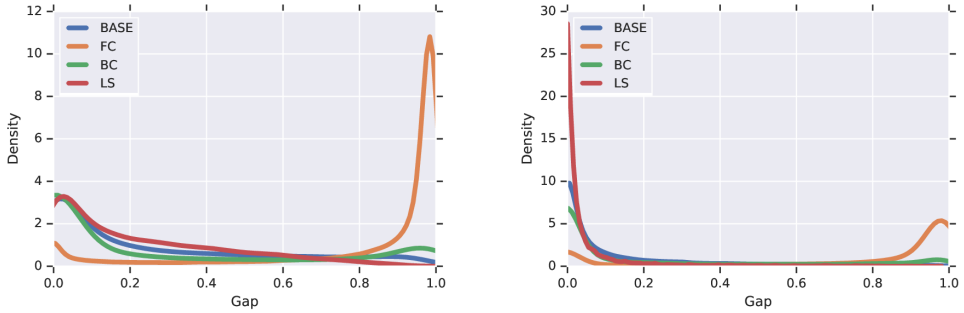
Model Calibration: To further tease out the impact of label smearing on model confidences, we ask: how do these techniques affect the *calibration* of the output probabilities? This measures how meaningful the model probabilities are in a frequentist sense (Dawid, 1982).

In Table 5, we report the expected calibration error (ECE) (Guo et al., 2017) on the test set for each method. While smoothing improves calibration over the baseline with $\alpha = 0.1$ — an effect noted also in (Müller et al., 2019) — for larger α, it becomes significantly *worse*. Furthermore, loss correction techniques significantly degrade calibration over smoothing. This is in keeping with the above findings as to these methods sharpening prediction confidences.



(a) CIFAR 100 (b) CIFAR 10

Figure 2. Effect of α on smoothing and forward label correction test accuracies on CIFAR-100 and CIFAR-10 from ResNet-32. Standard deviations are denoted by the shaded regions. Label smoothing (LS) significantly improves over baseline, and choosing $\alpha \gg \rho^*$, the true noise rate, improves even further. Forward correction (FC) outperforms LS and also benefits from choosing large values for α. Backward correction (BC) is worse than baseline for small α, and better than baseline for large α. In Table 8 in appendix, we report additional results for ResNet-56 and ResNet-32 from different label smearing methods, including where confusion matrix is estimated by pre-training a model as in Patrini et al. (2017).



(a) Gap between true (unobserved) label logit and mean logit.

(b) Gap between noisy (observed) label logit and mean logit.

Figure 3. Density of differences between logit corresponding to the true (left plot; corresponding to the "true" label, before injecting label noise) and noisy label (right plot; corresponding to the "noisy" label, after injecting label noise) and the average over all logits on the mis-labeled portion of the train data. Results are with $\alpha = 0.2$ on CIFAR-100, and the ResNet-32 model. LS reduces confidence mostly on the noisy label, whereas FC and BC increase confidence mostly on the true label. See Figure 7 for plots on full and clean data.

Summary: Overall, our results demonstrate that label smoothing is competitive with loss correction techniques in coping with label noise, and that it is particularly successful in denoising examples while preserving calibration.

4.2. Label Smoothing as Regularisation

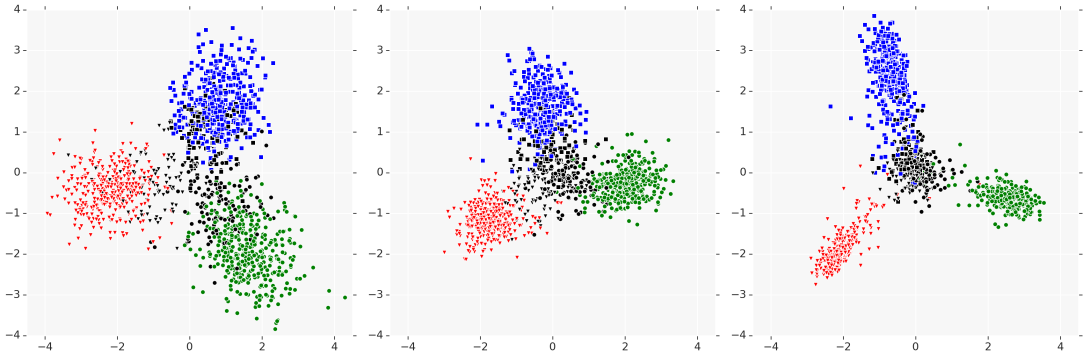
While empirically encouraging, the results in the previous section indicate a gap in our theoretical understanding: from §3.4, the smoothing loss apparently has the *opposite* effect to backward correction, which is theoretically unbiased under noise. What, then, explains the success of smoothing?

To understand the denoising effects of label smoothing, we now study its role as a *regulariser*. To get some intuition, consider a linear model $\mathbf{f}(x) = \mathbf{W}x$, trained on features $\mathbf{X} \in \mathbb{R}^{N \times D}$ and one-hot labels $\mathbf{Y} \in \{0,1\}^{N \times L}$ using the square loss, i.e., $\min_{\mathbf{W}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_2^2$. Label smoothing at level α transforms the optimal solution $\mathbf{W}^*$ to

$$\bar{\mathbf{W}}^* = (1-\alpha) \cdot \mathbf{W}^* + \frac{\alpha}{L} \cdot (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{J}. \quad (7)$$

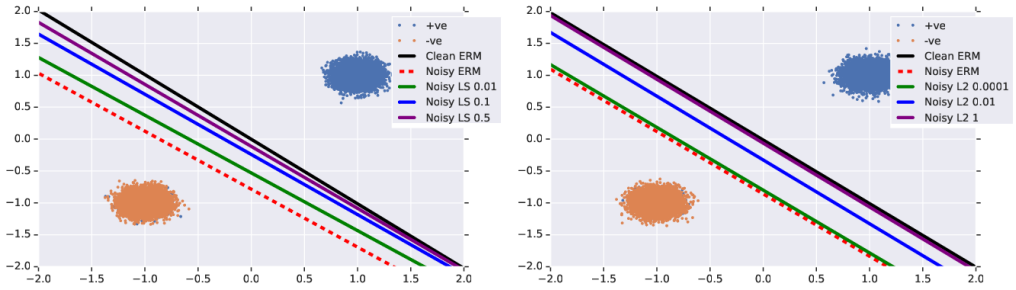
Observe that if our data is centered, the second term will be zero. Consequently, for such data, the effect of label smoothing is simply to shrink the weights. Thus, *label smoothing can have a similar effect to shrinkage regularisation.*

Our more general finding is the following. From (6), label smoothing is equivalent to minimising a *regularised* risk



(a) Label smoothing $\alpha = 0$. (b) Label smoothing $\alpha = 0.2$. (c) Label smoothing $\alpha = 0.7$.

Figure 4. Effect of label smoothing on pre-logits (penultimate layer output) under label noise. Here, we visualise the pre-logits of a ResNet-32 for three classes on CIFAR-100, using the procedure of Müller et al. (2019). The black markers denote instances which have been labeled incorrectly due to noise. Smoothing is seen to have a denoising effect: the noisy instances' pre-logits become more uniform, and so the model learns to not be overly confident in their label.



(a) Label smoothing. (b) ℓ_2 regulariser.

Figure 5. (a) Effect of label smoothing on logistic regression separator, on a synthetic problem with asymmetric label noise. The black line is the Bayes-optimal separator, found by logistic regression on the clean data. The other lines are separators learned by applying label smoothing with various α on the noisy data. Without smoothing, noise draws the separator towards the affected class; smoothing undoes this effect, and brings the separator back to the Bayes-optimal. (b) Shrinkage (ℓ_2) regularisation has a similar effect on the separator.

$R_{\text{sm}}(\mathbf{f}; D) \propto R(\mathbf{f}; D) + \beta \cdot \Omega(\mathbf{f})$, where

$$\Omega(\mathbf{f}) \doteq \mathop{\mathbb{E}}_{x} \left[\sum_{y' \in [L]} \ell(y', \mathbf{f}(x)) \right],$$

and $\beta \doteq \frac{\alpha}{(1-\alpha) \cdot L}$. The second term above does *not* depend on the underlying label distribution $\mathbb{P}(y \mid x)$. Consequently, it may be seen as a *data-dependent regulariser* on our predictor $\mathbf{f}$. Concretely, for the softmax cross-entropy,

$$\Omega(\mathbf{f}) = \mathop{\mathbb{E}}_{x} \left[L \cdot \log \left[\sum_{y'} e^{f_{y'}(x)} \right] - \sum_{y'} f_{y'}(x) \right]. \quad (8)$$

To understand the label smoothing regulariser (8) more closely, we study it for the special case of linear classifiers, i.e., $f_{y'}(x) = \langle \mathbf{W}_{y'}, x \rangle$. While we acknowledge that the label smoothing effects displayed in our experiments for deep networks are complex, as a first step, understanding these effects for simpler models will prove instructive.

Smoothing for Linear Models. For linear models $f_{y'}(x) = \langle \mathbf{W}_{y'}, x \rangle$, the label smoothing regularization for softmax cross-entropy (8) induces the following *shrinkage* effect.

Theorem 1. *Let x be distributed as $\mathbb{Q}$ with a finite mean. Then $\mathbf{W}_{y'} = 0, \forall y' \in [L]$ is a minimiser of (8). If further the data matrix $\mathbf{X}$ has rank $> L$, this is the unique minimiser.*

See Appendix A for the proof. We see that the label smoothing regulariser encourages shrinkage of our weights towards zero; this is akin to the observation for square loss in (7), and similar in effect to ℓ_2 regularisation, which is also motivated as increasing the classification margin.

This perspective gives one hint as to why smoothing may

Dataset	Architecture	Vanilla distillation	LS on teacher	LS on student	FC on teacher	FC on student
CIFAR-100	ResNet-32	63.98±0.26	64.48±0.25	63.83±0.28	66.65±0.18	63.94±0.34
CIFAR-100	ResNet-56	64.31±0.26	65.63±0.24	64.50±0.32	66.35±0.20	64.24±0.26
CIFAR-10	ResNet-32	80.44±0.64	86.95±1.82	85.72±2.61	86.81±1.86	86.92±2.11
CIFAR-10	ResNet-56	77.98±0.25	87.10±1.66	86.98±1.71	86.88±1.80	86.82±1.76

Table 6. Knowledge distillation experiments. We use label smoothing parameter $\alpha = 0.1$ and temperature parameter $T = 2$ during distillation, for all these experiments. We notice that doing LS on teacher improves the student accuracy compared to the baseline. LS on the student helps as well but not to the same accuracy. Loss correction using FC on teacher helps as well with the distillation.

successfully denoise. For linear models, introducing asymmetric label noise can move the decision boundary closer to a class. Hence, a regulariser that increases margin, such as shrinkage, can help the model to be more robust to noisy labels. We illustrate this effect with the following experiment.

Effect of Shrinkage on Label Noise. We consider a 2D problem comprising Gaussian class-conditionals, centered at $\pm(1, 1)$ and with isotropic covariance at scale $\sigma^2 = 0.01$. The optimal linear separator is one that passes through the origin, shown in Figure 5 as a black line. This separator is readily found by fitting logistic regression on this data.

We inject 5% *asymmetric* label noise into the negatives, so that some of these have their labels flipped to be positive. The effect of this noise is to move the logistic regression separator closer to the (true) negatives, indicating there is greater uncertainty in its predictions. However, if we apply label smoothing at various levels α, the separator is seen to gradually converge back to the Bayes-optimal; this is in keeping with the shrinkage property of the regulariser (8).

Further, as suggested by Theorem 1, an explicit L_2 regulariser has a similar effect to smoothing (Figure 5(b)). Formally establishing the relationship between label smoothing and shrinkage is an interesting open question.

Summary. We have seen in §3 that from a *loss* perspective, label smoothing results in a biased risk estimate; this is contrast to the unbiased backward correction procedure. In this section, we provided an alternate *regularisation* perspective, which gives insight into why label smoothing can denoise training labels. Combining these two views theoretically, however, remains an interesting topic for future work.

5. Distillation under Label Noise

We now study the effect of label smoothing on distillation, when our data is corrupted with label noise. In distillation, a trained "teacher" model's logits are used to augment (or replace) the one-hot labels used to train a "student" model (Hinton et al., 2015). While traditionally motivated as a means for a simpler model (student) to mimic the performance of a complex model (teacher), Furlanello et al. (2018) showed gains even for models of similar complexity.

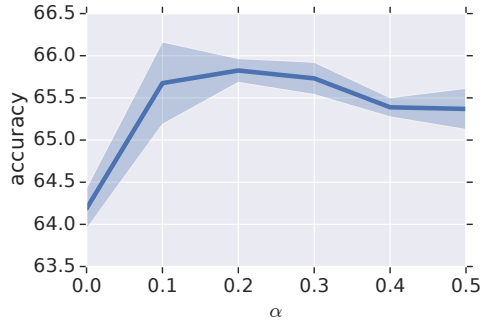


Figure 6. Effect of label smoothing on the teacher on student's accuracy after distillation with temperature $T = 1$, CIFAR-100. Teacher and student both use ResNet-32. For all values of α, label smoothing on the teacher improves distillation performance compared to a plain teacher ($\alpha = 0$).

Müller et al. (2019) observed that for standard (noise-free) problems, label smoothing on the teacher *improves* the teacher's performance, but *hurts* the student's performance. Thus, a better teacher does not result in a better student. Müller et al. (2019) attribute this to the erasure of relative information between the teacher logits under smoothing.

But is a teacher trained with label smoothing on *noisy* data better for distillation? On the one hand, as we saw in previous section, label smoothing has a denoising effect on models trained on noisy data. On the other hand, label smoothing on clean data may cause some information erasure in logits (Müller et al., 2019). Can the teacher transfer the denoising effects of label smoothing to a student?

We study this question empirically. On the CIFAR-100 and CIFAR-10 datasets, with the same architectures and noise injection procedure as the previous section, we train three teacher models on the noisy labels: one as-is on the noisy labels, one with label smoothing, and another with forward correction. We distill each teacher to a student model of the same complexity (see Appendix B for a complete description), and measure the student's performance. As a final approach, we distill a vanilla teacher, but apply label smoothing and forward correction on the *student*.

Table 6 reports the performance of the distilled students

using each of the above teachers. Our key finding is that on both datasets, *both label smoothing and loss correction on the teacher significantly improves over vanilla distillation*; this is in marked contrast to the findings of Müller et al. (2019). On the other hand, smoothing or correcting on the student has mixed results; while there are benefits on CIFAR-10, the larger CIFAR-100 sees essentially no gains.

Finally, we plot the effect of the teacher's label smoothing parameter α on student performance in Figure 6. Even for high values of α, smoothing improves performance over the baseline ($\alpha = 0$). Per the previous section, large values of α allow for successful label denoising, and the results indicate the value of this transfer to the student.

In summary, our experiments show that under label noise, it is strongly beneficial to denoise the teacher — either through label smoothing or loss correction — prior to distillation.

6. Conclusion

We studied the effectiveness of label smoothing as a means of coping with label noise. Empirically, we showed that smoothing is competitive with existing loss correction techniques, and that it exhibits strong denoising effects. Theoretically, we related smoothing to one of these correction techniques, and re-interpreted it as a form of regularisation. Further, we showed that when distilling models from noisy data, label smoothing of the teacher is beneficial. Overall, our results shed further light on the potential benefits of label smoothing, and suggest formal exploration of its denoising properties as an interesting topic for future work. More broadly, this work represents a step in studying the effects of common tricks in deep learning on label noise. Further expanding such study, as done for gradient clipping (Menon et al., 2020), is also of interest.

References

Amid, E., Warmuth, M. K. K., Anil, R., and Koren, T. Robust bi-tempered logistic loss based on bregman divergences. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 15013–15022. 2019.

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pp. 92–100, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. ACM.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 961–970, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Chorowski, J. and Jaitly, N. Towards better decoding and language model integration in sequence to sequence models. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pp. 523–527, 2017.

Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. Consistency of losses for learning from weak labels. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 197–210, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

Dawid, A. P. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Furlanello, T., Lipton, Z. C., Tschannen, M., Itti, L., and Anandkumar, A. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 1602–1611, 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017.

Han, B., Niu, G., Yao, J., Yu, X., Xu, M., Tsang, I. W., and Sugiyama, M. Pumpout: A meta approach for robustly training deep neural networks with noisy labels. *CoRR*, abs/1809.11008, 2018a.

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I. W.-H., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In *NeurIPS*, 2018b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., and Chen, z. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In Wallach, H., Larochelle, H., Beygelzimer, A., dÁlché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 103–112. Curran Associates, Inc., 2019.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1674–1684, Red Hook, NY, USA, 2017. Curran Associates Inc.

Li, W., Dasarathy, G., and Berisha, V. Regularization via structural label smoothing. *CoRR*, abs/2001.01900, 2020.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010. ISSN 0885-6125.

Menon, A. K., Rawat, A. S., Reddi, S. J., and Kumar, S. Can gradient clipping mitigate label noise? In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 4696–4705, 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. D., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1196–1204, 2013.

Patrini, G., Nielsen, F., Nock, R., and Carioni, M. Loss factorization, weakly supervised learning and label noise robustness. In *International Conference on Machine Learning (ICML)*, pp. 708–717, 2016.

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 *Computer Vision and Pattern Recognition (CVPR)*, pp. 2233–2241, 2017.

Pereyra, G., Tucker, G., Chorowski, J., Łukasz Kaiser, and Hinton, G. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations Workshop*, 2017.

Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized Evolution for Image Classifier Architecture Search. *arXiv e-prints*, February 2018.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping, 2014.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: consistency and maximal denoising. In *Conference on Learning Theory (COLT)*, pp. 489–511, 2013.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. In *ICLR Workshops*, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826, 2016.

Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. Combating label noise in deep learning using abstention. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6234–6243, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

van Rooyen, B. and Williamson, R. C. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

van Rooyen, B., Menon, A. K., and Williamson, R. C. Learning with symmetric label noise: the importance of being unhinged. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 10–18, 2015.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems 32*, pp. 6835–6846. Curran Associates, Inc., 2019.

Xie, L., Wang, J., Wei, Z., Wang, M., and Tian, Q. Disturblabel: Regularizing CNN on the loss layer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4753–4762, 2016.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004a.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5: 1225–1251, December 2004b. ISSN 1532-4435.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, June 2018.