Robustness and Reliability When Training With Noisy Labels

Amanda Olmin
Linköping University

Fredrik Lindsten
Linköping University

Abstract

Labelling of data for supervised learning can be costly and time-consuming and the risk of incorporating label noise in large data sets is imminent. When training a flexible discriminative model using a strictly proper loss, such noise will inevitably shift the solution towards the conditional distribution over noisy labels. Nevertheless, while deep neural networks have proven capable of fitting random labels, regularisation and the use of robust loss functions empirically mitigate the effects of label noise. However, such observations concern robustness in accuracy, which is insufficient if reliable uncertainty quantification is critical. We demonstrate this by analysing the properties of the conditional distribution over noisy labels for an input-dependent noise model. In addition, we evaluate the set of robust loss functions characterised by noise-insensitive, asymptotic risk minimisers. We find that strictly proper and robust loss functions both offer asymptotic robustness in accuracy, but neither guarantee that the final model is calibrated. Moreover, even with robust loss functions, overfitting is an issue in practice. With these results, we aim to explain observed robustness of common training practices, such as early stopping, to label noise. In addition, we aim to encourage the development of new noise-robust algorithms that not only preserve accuracy but that also ensure reliability.

1 INTRODUCTION AND PREVIEW

Deep neural networks have been successfully applied in many fields. However, because of their high complexity, they typically require a large amount of data for training. The process of annotating a large and possibly high-dimensional data set is costly, time-consuming and risks incorporating label noise in the training data set. For difficult boundary cases, even expert annotators can disagree about the true label. Evidently, as supervised training of discriminative models is highly dependent on the existence of labelled data, label noise poses a risk of hurting model performance. Firstly, the noise can cause the asymptotic risk minima to shift, resulting in a model approximating the conditional distribution over noisy, instead of clean, labels. Secondly, a flexible model might overfit to the noise in the data.

Central to the training of probabilistic predictive models are so called proper loss functions (Gneiting & Raftery, 2007; Reid & Williamson, 2010). A loss function is proper if it, for class variable Y and input X, is asymptotically minimised by the true conditional probability $f^*(x) = \mathbb{P}(Y \mid X = x)$. Moreover, if the (asymptotic) risk minimiser $\mathbb{P}(Y \mid X)$ is unique, the loss is *strictly proper*. Strictly proper loss functions, such as the commonly used categorical cross-entropy loss, encourage reliable uncertainty quantification during training. Using such a loss function, at least there is asymptotic, theoretical grounds for obtaining a model close to $\mathbb{P}(Y \mid X)$. A weaker, and perhaps more realistic, reliability condition than requiring that the model recovers the true class probability, is the notion of calibration. A model is calibrated if it reports class probabilities that agree with the observed prediction error frequency (see definition 2.2). Note that $\mathbb{P}(Y \mid X)$ is calibrated by definition.

In the presence of noisy labels, $\tilde{Y}$, the asymptotic risk minimiser of a strictly proper loss function is $\tilde{f}^*(x) = \mathbb{P}(\tilde{Y} \mid X = x)$ instead of $\mathbb{P}(Y \mid X = x)$. Hence, loss functions that are insensitive to noise have been proposed as alternatives, e.g. in (Charoenphakdee, Lee, & Sugiyama, 2019; Ghosh, Kumar, & Sastry, 2017;

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Ghosh, Manwani, & Sastry, 2015; Wang et al., 2019; Z. Zhang & Sabuncu, 2018). Specifically, fully *robust loss functions* are defined as having risk minimisers that are unaffected by label noise under certain assumptions on the noise distribution (Ghosh et al., 2017; Ghosh et al., 2015). Although such a condition is sufficient for achieving robustness in accuracy, it does not imply that the risk minimisers are reliable. This is a conceivable issue since uncertainty quantification is critical in many applications of machine learning. Intuitively, accurate reasoning about uncertainties becomes even more relevant in the case of noisy annotations.

In practice, the risk minimiser of a loss is only part of the story. As modern neural networks are often of high capacity, they are capable of overfitting to training data, also when labels are afflicted by noise (C. Zhang, Bengio, Hardt, Recht, & Vinyals, 2017). Conceptually, training with any loss function corresponds to a, possibly implicit, assumption that the training trajectory will pass "close" to the risk minimiser f^*, before drifting off into overfitting. We illustrate this in fig. 1a where we think of the training dynamics as consisting of two phases: a convergent phase where the model approaches f^*, followed by a divergent (overfitting) phase where the distance between the model and the risk minimiser increases. We will not elaborate on the details of the training dynamics, nor characterise what "close" means. Still, we argue that this assumption underlies common practices of training neural networks, such as using early stopping to halt the trajectory as close as possible to f^*. It is further supported by observations that overparameterised models tend to learn general patterns in the training data before overfitting to noisy examples (Arpit et al., 2017). We will use this assumption of two phases of training to illustrate some of the key points in the paper.

In the view of fig. 1b, using a strictly proper loss function, the convergent phase of training will "aim" towards $f^*(x) = \mathbb{P}(Y \mid X = x)$ in the case of noise-free data and towards $\tilde{f}^*(x) = \mathbb{P}(\tilde{Y} \mid X = x)$ if labels are noisy. As a first contribution of this paper we therefore:

Characterise the Properties of $\mathbb{P}(\tilde{\mathbf{Y}} \mid \mathbf{X})$ Relative to $\mathbb{P}(\mathbf{Y} \mid \mathbf{X})$. We show (proposition 3.1) that under the commonly used symmetric label noise assumption, as well as for a more realistic input-dependent noise, $\tilde{f}^*(x) = \mathbb{P}(\tilde{Y} \mid X = x)$ shares decision boundaries with $f^*(x) = \mathbb{P}(Y \mid X = x)$. This is encouraging if accuracy is the main quantity of interest, and can explain observed robustness in accuracy of neural networks trained with regularisation techniques such as early stopping (Li, Soltanolkotabi, & Oymak, 2020) and pre-training (Hendrycks, Lee, & Mazeika, 2019). Still, for reliable uncertainty quantification, this is not enough. Considering the two risk minimisers, we show (proposition 3.2) that $\tilde{f}^*$ has higher entropy than f^*. While this is perhaps not surprising, since the additional noise can only increase the entropy, it still shows that we will not be able to recover the true probability $\mathbb{P}(Y \mid X)$, despite the apparent robustness in accuracy. Of a higher interest, and with more severe practical implications, is that the noisy risk minimiser $\tilde{f}^*$ is not only insufficient for recovering $\mathbb{P}(Y \mid X)$, but it also fails to be calibrated (proposition 3.3). We demonstrate in fig. 2 how the predicted class probability of a simple neural network trained with early stopping is affected by the presence of symmetric label noise in the training data. The models trained with and without label noise have similar accuracy on clean test data (0.984 vs. 0.992), but the uncertainty clearly increases with the addition of the noise.

Building on these findings, as a second contribution we also:

Critically Review the Use of Robust Loss Functions. Employing a robust loss function (Ghosh et al., 2017; Ghosh et al., 2015), means that $\tilde{f}^* = f^*$, i.e., as illustrated in fig. 1c, the training trajectories "aim" for the same point in the initial training phase. However, we argue that this is not enough when it comes to reliability. Indeed, we show (proposition 3.4) that robust loss functions are never strictly proper, so we can not expect them to accurately recover $\mathbb{P}(Y \mid X)$. Furthermore, to relax this strong requirement, we define a weaker notion of a calibration-based strictly proper loss function (definition 3.1), and show (proposition 3.5) that the robustness condition is insufficient for a loss function to be calibration-based strictly proper. Specifically, the set of symmetric, robust loss functions (Ghosh et al., 2017; Ghosh et al., 2015) are never calibration-based strictly proper. This is of high relevance since, to our knowledge, this is the only identified class of loss functions that are robust to simple non-uniform label noise and that also does not require estimation of the noise distribution.

Finally, we demonstrate empirically (section 3.3) that models trained with robust loss functions are *not* robust to overfitting and that any observed robustness does not follow from the theory. Indeed, robustness, in this regard, is a property related to the asymptotic risk minimiser f^*. In practice, a model can overfit to the noise in the training data even when a robust loss function is used.

In summary, loss robustness concerns the risk minimiser f^* which determines the "aim" of the convergent phase of the learning trajectory. However, our results show that there is limited theoretical support for why this target point should be any better using a robust (fig. 1c) compared to a strictly proper loss function (fig. 1b). Specifically, in the case of a symmetric, robust loss

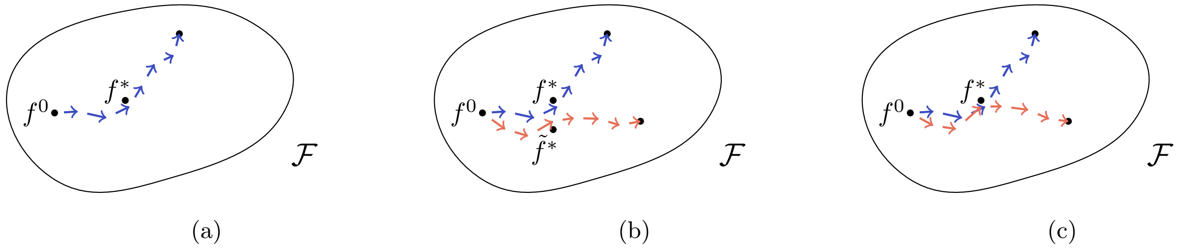


Figure 1: (a) Sketch of trajectory in model space $\mathcal{F}$ during training. In the convergent phase, the trajectory "aims" towards the risk minimiser f^*. In the divergent (overfitting) phase, the distance between the model and the risk minimiser increases. (b) When labels are noisy, the trajectory aims towards the noisy risk minimiser $\tilde{f}^*$. Using a strictly proper loss, the clean and noisy risk minimisers differ. (c) Using a robust loss, the aim of the convergent phase is the same under the clean and noisy data distributions, but overfitting is still possible.

functions, it holds that both are robust in terms of accuracy, none of them are robust when it comes to uncertainty quantification (whether we use the stronger notion of recovering $\mathbb{P}(Y \mid X)$ or the weaker notion of calibration), and the practical issue of potentially overfitting to label noise remains in both cases.

Furthermore, while our results can be used to explain perceived robustness of specific training algorithms, they also point towards a weakness in evaluating robustness solely in terms of accuracy. They demonstrate that robustness in accuracy does not imply reliability, or, more specifically, robustness in uncertainty quantification. At the same time, uncertainty-agnostic metrices such as accuracy are commonly used to evaluate robustness against label noise (Song, Kim, Park, & Lee, 2020) while, to our understanding, uncertainty quantification is consistently overlooked. Our conclusion is that further investigation is needed to better understand the effect of label noise on model reliability, as well as the training dynamics. Moreover, we suggest that uncertainty quantification should have a natural part in the evaluation of noise-robust training algorithms. For future work, one potential direction is that of developing loss functions that ensure robustness in accuracy as well as calibration-based strictly properness.

2 PRELIMINARIES

We will consider classification problems with K classes, input variable $X \in \mathcal{X}$ and label $Y \in \mathcal{Y} = \{1, \ldots, K\}$. We will denote the true vector of probabilities over outcomes in $\mathcal{Y}$ given the observation $X = x$ by $g(x)$ with elements $g_k(x) = \mathbb{P}(Y = k \mid X = x)$. When there is label noise in the data, we observe $\tilde{Y} \in \mathcal{Y}$ in place of Y and $\tilde{g}(x)$ is used to denote the probability vector with entries $\tilde{g}_k(x) = \mathbb{P}(\tilde{Y} = k \mid X = x)$.

For the generative process of $\tilde{Y}$, we will consider a version of input-dependent noise referred to as *simple non-uniform* label noise (Ghosh et al., 2017).

Definition 2.1 (Simple non-uniform label noise, see e.g. Ghosh et al., 2017)**.** *For simple non-uniform label noise,*

$$\mathbb{P}(\tilde{Y} = \tilde{y} \mid Y = y, X = x) = \begin{cases} 1 - \omega(x), & \textit{if } \tilde{y} = y \\ \frac{\omega(x)}{K-1}, & \textit{otherwise} \end{cases}$$

where the flip probabilities are defined by an input-dependent noise parameter $0 \leq \omega(x) < \frac{K-1}{K}$.

In the literature, it is common to assume that label noise is input-independent, see (Song et al., 2020) and the references therein. Hence, although referred to as "simple" the simple non-uniform label noise assumption is still more complex than what is often assumed, since it allows the noise to vary across the input space. In parallel, this noise assumption does not exclude the possibility of input-independence. For instance, the frequently used *symmetric* label noise (Song et al., 2020), for which $\omega(x) = \omega \; \forall x$, is a special case of simple non-uniform label noise.

The bounds on $\omega(x)$ given in definition 2.1 will be implicitly assumed throughout the paper. The upper bound on the noise parameter $\omega(x)$ ensures that the noisy label has a higher probability of being equal to the true label than it has of being equal to any other label. Theoretically, this will preserve the dominant label in each cluster of a sampled data set. Throughout the paper, we will assume that $\tilde{g}(X) \neq g(X)$ with probability larger than 0. For simple non-uniform label noise, this is equivalent to assuming that $\mathbb{P}\left(\{\omega(X) > 0\} \cap \{g(X) \neq \frac{1}{K} \cdot \mathbf{1}_K\}\right) > 0$, where $\mathbf{1}_K$ is the vector of ones of size K. For the complement, the label noise is effectively non-existent.

2.1 Risk Minimisation and Reliability

The aim is to train a model $f : \mathcal{X} \to \Delta^{K-1}$, belonging to some model class $\mathcal{F}$ and predicting a conditional

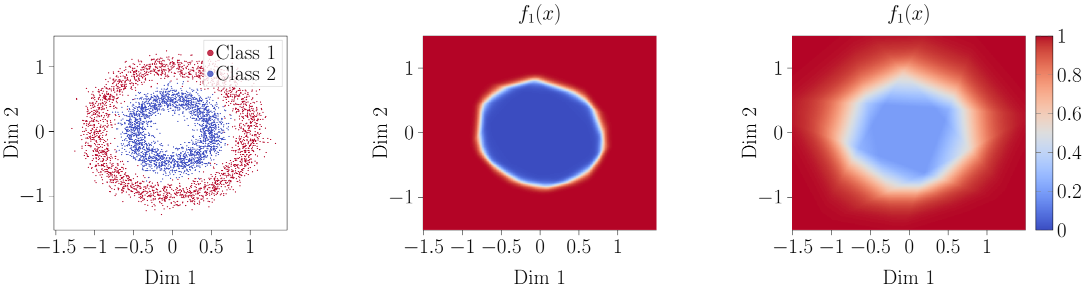


Figure 2: Predicted class 1 probabilities, $f_1(x)$, over the input space, of models trained on circle data. Left: clean training data (Pedregosa et al., 2011). Middle: model trained on clean data. Right: model trained on noisy data with label flip probability 0.2.

distribution over Y for each $x \in \mathcal{X}$. Here, Δ^{K-1} denotes the $(K-1)$-simplex and so, $f_k(x) \in [0,1]\ \forall k$ and $\|f(x)\|_1 = \sum_{k\in\mathcal{Y}} f_k(x) = 1$.

We will consider risk minimisation, where the risk with respect to the data distribution $\mathbb{P}(X,Y)$ is defined according to

$$\mathcal{R}_\ell(f) = \mathbb{E}_X[\mathbb{E}_{Y|X}[\ell(f(X),Y)]] \qquad (1)$$

with $\ell(f(X),Y)$ a predefined loss function. In practice, the true data distribution is unknown and the risk is approximated by the empirical risk with respect to a finite data set.

Optimally, we would minimise $\mathcal{R}_\ell$ directly. However, when noisy labels are observed, minimising the noisy risk $\tilde{\mathcal{R}}_\ell$, with respect to the noisy data distribution $\mathbb{P}(X,\tilde{Y})$, is arguably the most straightforward, and occasionally the most sensible, option. In this case, the aim is nevertheless to find a model that performs well on the intended, clean target distribution. Hence, we will consider the properties of the model obtained by minimising $\tilde{\mathcal{R}}_\ell$ and its relation to Y.

Central to the arguments put forth in this paper, is the idea that reliability is a model property that can be equally important as accuracy. Formally, reliability (or calibration) is defined as follows.

Definition 2.2 (Calibrated model, see e.g. Bröcker, 2009; Vaicenavicius et al., 2019)**.** *Let $f : \mathcal{X} \to \Delta^{K-1}$ be a probabilistic predictive model and assume that $\mathbb{P}(Y \mid f(X))$ exists. The model f is calibrated if*

$$\mathbb{P}(Y \mid f(X)) = f(X)$$

almost surely.

Hence, a model is reliable if its confidence, represented by the predicted class probabilities, is equal to the true conditional probability over the outcome.

2.2 Proper and Robust Loss Functions

In risk minimisation, the use of a strictly proper loss function gives asymptotic, theoretical guarantees that the true conditional distribution over labels will be recovered at the minimum.

Definition 2.3 (Proper loss function, see e.g. Gneiting and Raftery, 2007; Reid and Williamson, 2010)**.** *A loss function ℓ is proper if*

$$g(x) \in \underset{f(x)\in\Delta^{K-1}}{\operatorname{argmin}}\ \mathbb{E}_{Y|X=x}[\ell(f(x),Y)]$$

for all conditional distributions $\mathbb{P}(Y \mid X)$. If the minimum at $g(x)$ is unique, ℓ is strictly proper.

Definition 2.3 concerns the point-wise risk of proper loss functions. However, by minimising the point-wise risk, we implicitly minimise the full risk. We will denote the set of proper and strictly proper loss functions by $\mathcal{L}_{\mathrm{P}}$ and $\mathcal{L}_{\mathrm{SP}}$, respectively.

Under the presence of label noise, the asymptotic risk minimiser of a strictly proper loss function will be $\tilde{g}$ as opposed to g. Hence, the risk minimiser will differ depending on if Y or $\tilde{Y}$ is considered. In contrast to this, robust loss functions have been identified and developed based on the idea of achieving robustness through noise-insensitive risk minimisers (e.g. (Ghosh et al., 2017; Ghosh et al., 2015; Wang et al., 2019; Z. Zhang & Sabuncu, 2018)). In this context, a loss ℓ is said to be robust to label noise if the asymptotic risk minimisers of the clean and noisy risks have the same probability of misclassification. A sufficient condition is that the risk minimiser, f^*, of the clean risk, is also a minimiser of the noisy risk (Ghosh et al., 2017).

Definition 2.4 (Robust loss function, see Ghosh et al., 2017; Ghosh et al., 2015)**.** *A loss function ℓ is robust to label noise if for all asymptotic minimisers f^* of the clean risk, $\mathcal{R}_\ell$, it holds that*

$$\tilde{\mathcal{R}}_\ell(f^*) \le \tilde{\mathcal{R}}_\ell(f), \quad \forall f \in \mathcal{F}$$

where $\tilde{\mathcal{R}}_\ell$ is the risk under the noisy data distribution.

We will refer to the set of robust loss functions by $\mathcal{L}_\mathrm{R}$. Definition 2.4 gives a theoretical condition for robustness, but does not, in itself, tell us how to construct a robust loss function. Hence, there is a practical need of identifying individual loss functions, or classes thereof, that fulfills this condition. Ghosh et al., 2017; Ghosh et al., 2015, find that *symmetric* loss functions are robust under symmetric label noise and under simple non-uniform label noise with the extra condition that ℓ is positive and $\mathcal{R}_\ell(f^*) = 0$.

Definition 2.5 (Symmetric loss function, see e.g. Ghosh et al., 2017; Ghosh et al., 2015)**.** *A loss function ℓ is symmetric if*

$$\sum_{k=1}^{K} \ell(q, k) = C, \quad \forall q \in \Delta^{K-1}$$

for some constant C.

Referring to the set of symmetric loss functions by $\mathcal{L}_\mathrm{S}$, it thus holds that $\mathcal{L}_\mathrm{S} \subseteq \mathcal{L}_\mathrm{R}$. To our knowledge, $\mathcal{L}_\mathrm{S}$ is the only identified set of loss functions that are robust to simple non-uniform label noise and that does not require an estimate of $\omega(x)$. We recognise that there exists other robust loss functions but that is either only robust to input-independent label noise (Xu, Cao, Kong, & Wang, 2019) or rely on knowledge of $\omega(x)$ (see e.g. Natarajan, Dhillon, Ravikumar, and Tewari, 2018; Patrini, Rozza, Menon, Nock, and Qu, 2017). We refer to the supplementary material for an analysis of the information-theoretic loss function proposed in (Xu et al., 2019). For the second group of loss functions, we do not consider them in this paper, with the argument that the true noise rates are seldomly known in practice.

3 MAIN RESULTS

We evaluate the use of strictly proper and robust loss functions in the presence of label noise by analysing their respective risk minimisers. In general, we consider the following asymptotic properties of a loss function under the influence of label noise:

(A) It preserves decision boundaries, or accuracy, of a model trained with clean data.

(B) It recovers the true conditional probability $g(x)$.

(C) It results in a reliable, or calibrated, model.

We find that both strictly proper and robust loss functions fulfill (A), but neither gives asymptotic, theoretical guarantees for (B) or (C). Moreover, strictly proper as well as robust loss functions are susceptible to overfitting in practice. In parallel, while (A) is commonly considered in the context of label noise robustness, (B) and (C) are consistently overlooked. This, we argue, in spite of their equal importance in many practical applications. The complete proofs and derivations for the results presented in this section can be found in the supplementary material.

3.1 Distribution Over Noisy Labels

Using a strictly proper loss function, we expect a flexible model to asymptotically approximate g under the clean data distribution and $\tilde{g}$ if label noise is present in the data. Hence, we expand on the implications of training a model with a data set containing simple non-uniform label noise by deriving results regarding the conditional distribution over noisy labels. To evaluate $\tilde{g}$, we first note that it can be derived from g by marginalisation

$$\tilde{g}(x) = \sum_{k=1}^{K} \mathbb{P}(\tilde{Y} \mid Y = k, X = x) g_k(x). \tag{2}$$

An alternative formulation is introduced in the following lemma.

Lemma 3.1. *The conditional probability vector $\tilde{g}(x)$ can be written as function of $g(x)$ according to*

$$\tilde{g}(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) g(x) + \frac{\omega(x)}{K-1} \cdot \mathbf{1}_K$$

where $\mathbf{1}_K$ is the vector of ones with length K. Moreover, it holds for any two classes $i, j \in \mathcal{Y}$ that $\tilde{g}_i(x) > \tilde{g}_j(x)$ if and only if $g_i(x) > g_j(x)$.

From lemma 3.1, the first result regarding the properties of $\tilde{g}$ follows.

Proposition 3.1. *Assume that the prediction is taken as the most probable class, then $\tilde{g}$ has the same decision boundaries as g.*

Proposition 3.1 establishes that the use of a strictly proper loss function will asymptotically preserve the accuracy of g in the presence of simple non-uniform label noise. It is noteworthy, that if g and $\tilde{g}$ had not shared decision boundaries, no classification-calibrated loss function would be robust to the noise in terms of accuracy. Simply put, a loss function is classification-calibrated if the class predictions of its asymptotic risk minimiser(s) corresponds to taking the most probable class with respect to the true conditional probability of the observed target variable (Bartlett, Jordan, & McAuliffe, 2006). In binary classification, it is a minimal condition commonly imposed on surrogate losses of the 0/1-risk.

Next, we will show that preserving accuracy is not enough if uncertainty quantification is critical. As a first step towards this realisation, we consider the entropy of $\tilde{g}$. The conditional entropy of a probabilistic vector $f(x)$ given $X = x$ is defined as

$$\mathcal{H}[f(x)] = -\sum_{k=1}^{K} f_k(x) \log f_k(x). \tag{3}$$

We have the following result.

Proposition 3.2. *The average conditional entropy of $\tilde{g}$ is higher than that of g, that is $\mathbb{E}_X[\mathcal{H}[\tilde{g}(X)]] > \mathbb{E}_X[\mathcal{H}[g(X)]]$.*

Since $\tilde{g}(X) \neq g(X)$ with non-zero probability, proposition 3.2 implies that training with a strictly proper loss function, under the influence of label noise, will not recover the true conditional probability over clean labels. This does not necessarily entail that $\tilde{g}$ is uncalibrated. Nevertheless, the next result states that it is.

Proposition 3.3. *The vector of conditional probabilities $\tilde{g}(X)$ is not calibrated with respect to the distribution $\mathbb{P}(Y \mid X)$ over clean labels.*

We conclude that proposition 3.1 is in agreement with the observed robustness to label noise in accuracy when training with regularisation (Hendrycks et al., 2019; Li et al., 2020). If the model does not overfit to the data, such that it approximates $\tilde{g}$, it will have similar accuracy on noise-free data as a model approximating g. However, propositions 3.2 and 3.3 imply that this does not generalise to uncertainty quantification. Training with noise in the data using a strictly proper loss, will not asymptotically recover g nor result in a reliable model. Hence, evaluating robustness solely in terms of accuracy can give a perceived robustness against label noise in spite of the final model not being reliable.

3.2 Evaluation of Robust Loss Functions

We derive results concerning the set of robust loss functions in definition 2.4. Robustness in this regard has previously been determined to be sufficient for preserving accuracy (Ghosh et al., 2017; Ghosh et al., 2015). Hence, our analysis concentrates on the reliability of the asymptotic risk minimisers of robust loss functions, shared between the clean and noisy risks. We first establish that robust loss functions are not strictly proper.

Proposition 3.4. *Robust loss functions (definition 2.4) are not strictly proper.*

Intuitively, if a robust loss function had been strictly proper, it should have had a unique minimum at g under the clean data distribution and at $\tilde{g}$ under the noisy data distribution and hence, there would be no overlap in risk minimisers unless $\tilde{g} = g$. The opposite of proposition 3.4 must also be true; strictly proper loss functions are not robust according to definition 2.4 (a proof for $K = 2$ can be found in (Reid & Williamson, 2010)). Thus, $\mathcal{L}_{\mathrm{SP}} \cap \mathcal{L}_{\mathrm{R}} = \emptyset$.

Next, we consider the reliability of the asymptotic risk minimisers of robust loss functions. First, we again point out that, to our knowledge, the only identified class of loss functions that fulfills definition 2.4 under simple non-uniform label noise and at the same time does not require an estimate of the noise parameter $\omega(x)$ is that of symmetric loss functions (Ghosh et al., 2017; Ghosh et al., 2015). Hence, the following results focus on this class.

We have shown that, in the context of label noise, using a strictly proper loss function is not optimal if reliability is of importance. Hence, it is of higher relevance to investigate if a robust loss function can recover g, than if it is strictly proper. We argue, however, that this is not an inherent property of robust loss functions in general. The conditional (or point-wise) risk minima $f^*(x) = [f_1^*(x), 1 - f_1^*(x)]^\top$ for a symmetric loss function, in a binary classification setting, are found at

$$f_1^*(x) = \gamma \mathbb{1}_{\mathbb{P}(Y=1|X=x) \geq \frac{1}{2}} + \gamma' \mathbb{1}_{\mathbb{P}(Y=1|X=x) < \frac{1}{2}} \tag{4}$$

with $\gamma \in \operatorname{argmin}_{q \in [0,1]} \ell(q, 1)$ and $\gamma' \in \operatorname{argmax}_{q \in [0,1]} \ell(q, 1)$. From here on, we will assume that $\gamma \in [\frac{1}{2}, 1]$ and $\gamma' \in [0, \frac{1}{2})$, since in all other cases for which $f_1^*(x) \in [0, 1]$, predictions will be consistently incorrect for at least one class. In agreement with (Charoenphakdee et al., 2019), we can not mathematically recover the true conditional probability over Y from the minima in eq. (4). As a result, g can not be a unique minimiser of the clean (or noisy) risk of a symmetric loss function in general. Consequently, definition 2.4 is not sufficient for recovering g.

While an arbitrary robust loss function will not asymptotically recover g, this does not exclude the possibility of obtaining a calibrated model. Unfortunately, we demonstrate that the two properties of fulfilling the robustness condition in definition 2.4 and having only calibrated risk minimisers do not coincide. To this end, we will introduce a new set of loss functions referred to as *calibration-based strictly proper*. We denote this set of loss functions by $\mathcal{L}_{\mathrm{CSP}}$.

Definition 3.1. *(Calibration-based strictly proper loss function) Let $\mathcal{F}_\mathcal{C}$ be the set of calibrated models in $\mathcal{F}$. The loss function ℓ, with asymptotic risk minimisers $f^* \in \mathcal{F}$, is calibration-based strictly proper if*

$$f^* \in \mathcal{F}_\mathcal{C}, \quad \forall f^* \in \mathcal{F},$$

for all $\mathbb{P}(Y \mid X)$ and for all input distributions μ_X.

In parallel to the definition of a strictly proper loss function, the definition of a calibration-based strictly proper loss puts a restriction on the asymptotic minimiser(s) of the corresponding risk. However, instead of requiring that the (unique) risk minimiser is equal to g, it requires that all asymptotic risk minimisers are calibrated. Hence, employing $\ell \in \mathcal{L}_{\text{CSP}}$ will, asymptotically, result in a calibrated model. Since g is calibrated by definition, all strictly proper loss functions are calibration-based strictly proper. However, this is not true for symmetric loss functions.

Proposition 3.5. *Symmetric loss functions (definition 2.5) are not calibration-based strictly proper.*

From proposition 3.5, we have $\mathcal{L}_{\text{S}} \cap \mathcal{L}_{\text{CSP}} = \emptyset$. As an example, mean absolute error and sigmoid loss, both symmetric, have a unique minimum for $f_1(x) \in [0, 1]$ in eq. (4) with $(\gamma, \gamma') = (1, 0)\ \forall x \in \mathcal{X}$. For a non-separable classification problem, this model is never calibrated. Nevertheless, symmetric loss functions are robust to symmetric label noise also in this case.

To conclude, definition 2.4 constitutes a sufficient condition for robustness in accuracy. However, there are no theoretical basis for why, even asymptotically, the use of a robust loss function will result in a reliable model. On the contrary, the class of symmetric, robust loss functions will *not* (asymptotically) recover g and are *not* calibration-based strictly proper.

3.3 Robustness and Overfitting

While definition 2.4 relies on asymptotic theory, a finite training set has to suffice in practice. Naturally, empirical evaluation has been used to demonstrate the noise-insensitivity and motivate the use of robust loss functions also under such circumstances (Ghosh et al., 2017). However, we show, with a similar experiment, that these loss functions are not robust to overfitting and argue that any perceived robustness is not explained by the asymptotic theory.[1]

For the experiment, we consider neural networks with one hidden layer of 500 hidden units and with LeakyReLU as activation function. The data used is flattened MNIST images (LeCun, Bottou, Bengio, & Haffner, 1998) to which we artificially add symmetric label noise with parameter $\omega \in [0.0, 0.3, 0.5]$. We train the models using ADAM optimization (Kingma & Lei Ba, 2015) with a constant learning rate of 0.005 and a batch size of 100. Similar to Ghosh et al., 2017, the robust (and symmetric) loss function that we consider is mean absolute error (MAE). First, we train models with MAE and random initialisation. We evaluate the accuracy on the corresponding training data set and a noise-free test data set during the course of training, as shown in fig. 3b. For comparison, we do a similar evaluation of models trained with categorical cross-entropy (CCE) loss, see fig. 3a.

Under the influence of label noise, the models trained with CCE clearly overfit to the training data. The accuracy on the training data set is close to 1 for all considered values of ω within 500 epochs of training, while the test accuracy decreases as training progresses. In contrast, the models trained with MAE seems to stabilise at a point where the accuracy over the clean test data set remains high, even in the presence of label noise. Observing only these trends, it is intriguing to assume that MAE is robust to overfitting and that this can be explained by the loss function's noise-insensitive risk minimiser. However, overfitting is not considered in asymptotic theory, on which the noise-insensitivity of robust loss functions rely. In addition, we have previously demonstrated that the asymptotic risk minimiser of CCE loss, a strictly proper loss, is also robust in accuracy to symmetric label noise. In spite of this, the models trained with CCE loss overfit to the label noise.

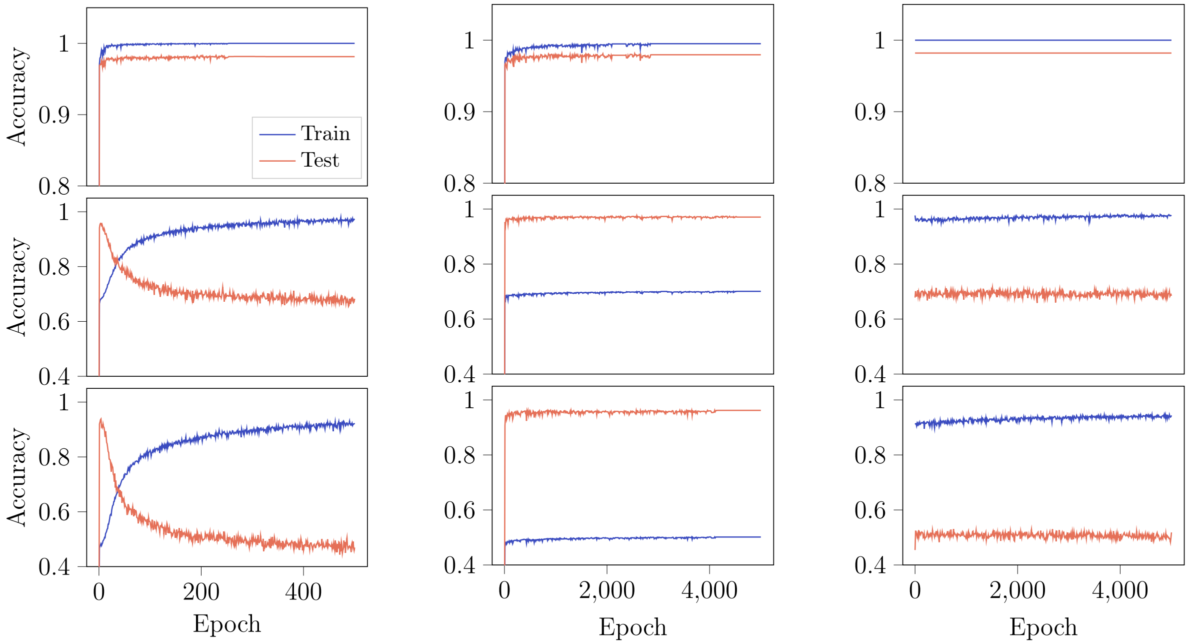
To support our arguments, we train models again with MAE but replace the random initialisation. For each noise level, the model weights are initialised with those obtained when training a model with CCE loss for 500 epochs. Evaluating the new models in terms of accuracy, the trends observed are more similar to those of the models trained only with CCE loss, see fig. 3c. Evidently, the models are capable of overfitting to training data, even when a robust loss function is used. Furthermore, in all cases, the models achieve a smaller training loss compared to those trained with random initialisation, as shown in fig. 4. Hence, the models trained with random initialisation must be stuck in local minima. At the same time, the asymptotic theory of robust loss functions concerns global, not local, minima.

To conclude, we have demonstrated that robustness in the context of definition 2.4 should not be confused with robustness to overfitting. Indeed, the asymptotic theory behind robust loss function is not concerned with this issue. Thus, models can overfit to label noise, even when a robust loss function is employed.

4 DISCUSSION

On one hand, the results presented in this paper can help explain why some training algorithms have a perceived, inherent robustness to label noise and are re-

[1] Code provided at: https://github.com/AOlmin/robustness_and_reliability_in_weak_supervision



(a) CCE

(b) MAE

(c) MAE, pre-trained

Figure 3: Accuracy on (noisy) train and clean test MNIST data for models trained with symmetric label noise. Models are trained with categorical cross-entropy loss (CCE) or mean absolute error (MAE). Models trained with MAE are either initialised randomly (MAE) or pre-trained with CCE loss (MAE, pre-trained). Top: $\omega = 0.0$. Middle: $\omega = 0.3$. Bottom: $\omega = 0.5$.

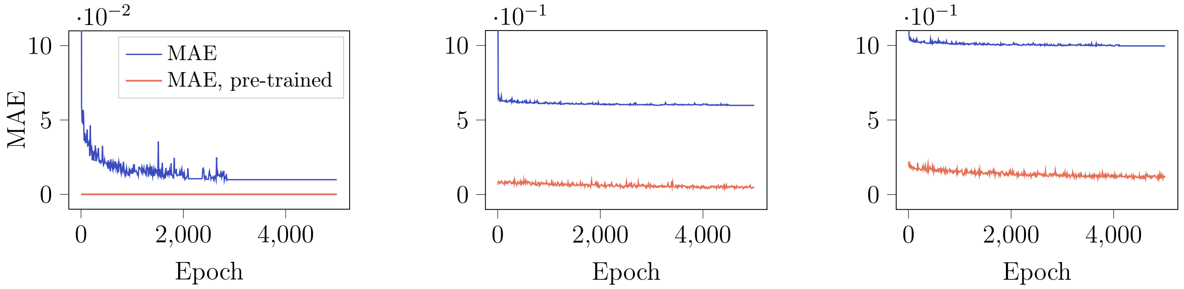
assuring if accuracy is the main property of interest. On the other hand, they point towards a weakness of evaluating robustness solely based on metrices that are agnostic to predicted class probabilities. Specifically, it does not ensure that the final model is reliable. Although we have considered a simpler input-dependent noise model in this paper, the hope is that the knowledge gained could be used as a basis for gaining understanding of more complex noise. Moreover, part of the results presented are of the negative kind. Hence, if they do hold for the specific noise model considered, they also hold in the more general case.

Consider training a flexible discriminative model, such as a deep neural network, capable of approximating the true conditional probability over labels. When labels are noisy and using a strictly proper loss function, the convergent phase of the training dynamics will "aim" towards $\tilde{f}^* = \tilde{g}$ instead of $f^* = g$, as illustrated in fig. 1b. We have shown that, in this case, $\tilde{f}^*$ share decision boundaries with f^*, but it is not calibrated. From the view of fig. 1c, using a robust loss function, the convergent phase of the training trajectory will instead aim towards the same point, regardless of whether labels are noisy or not, i.e. $\tilde{f}^* = f^*$. We have characterised the properties of this risk minimiser and conclude that fulfillment of the robustness condition (definition 2.4) is neither a proxy for having $\tilde{f}^* = g$ nor for $\tilde{f}^*$ being calibrated.

With these results in mind, we argue that there is *no theoretical motivation* for why an arbitrary robust loss function would be better to employ than a strictly proper loss function. Under simple non-uniform noise, both are robust in terms of accuracy, but neither asymptotically guarantees that g will be recovered or that the final model will be calibrated. In section 3.2, we introduced the class of calibration-based strictly proper loss functions, for which all asymptotic risk minimisers are calibrated. Our argument is that there is a relevance in defining a loss function that, asymptotically, achieves robustness in accuracy as well as results in a calibrated model. For future work, it would be of interest to find a class of loss functions that fulfills both of these conditions and that is also not dependent on an estimate of the (usually unknown) noise distribution.

5 CONCLUSION

For supervised training of discriminative models, we investigated the effect of label noise on model performance by analysing the properties of the conditional distribution over noisy labels. Furthermore, we critically reviewed the set of robust loss functions characterised by asymptotic risk minimisers that are insensitive to



(a) $\omega = 0.0$

(b) $\omega = 0.3$

(c) $\omega = 0.5$

Figure 4: Train loss of models trained with MAE. The models pre-trained with CCE achieve a smaller loss than the models trained with random weight initialisation.

label noise. Under an input-dependent noise model, we found that both strictly proper and robust loss functions offer asymptotic robustness in accuracy but neither offer asymptotic, theoretical guarantees for obtaining a calibrated, or reliable, model in the presence of label noise. In addition, robustness in this context should not be misunderstood as the model being robust to noise in practice. Even when a robust loss function is used, the model can still overfit to training data. We conclude that further investigation is needed for better understanding of the effects of label noise on model performance and in order to ensure reliability of models trained with label noise. Such research would be valuable in the strive for safe employment in society.

Acknowledgements

We thank Jakob Lindqvist and Lennart Svensson (Chalmers University of Technology, Sweden) for constructive feedback on the paper.

This research is financially supported by the Swedish Research Council via the project *Handling Uncertainty in Machine Learning Systems* (contract number: 2020-04122), the Swedish Foundation for Strategic Research via the project *Probabilistic Modeling and Inference for Machine Learning* (contract number: ICA16-0015), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and ELLIIT.

References

Arpit, D., Jastrzębskl, S., Bailas, N., Krueger, D., Bengio, E., Kanwal, M. S., … Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In *International Conference on Machine Learning* (pp. 350–359).

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

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(643), 1512–1519.

Charoenphakdee, N., Lee, J., & Sugiyama, M. (2019). On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning* (pp. 961–970).

Ghosh, A., Kumar, H., & Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence* (pp. 1919–1925).

Ghosh, A., Manwani, N., & Sastry, P. S. (2015). Making risk minimization tolerant to label noise. *Neurocomputing*, *160*, 93–107.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Hendrycks, D., Lee, K., & Mazeika, M. (2019). Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning* (pp. 4815–4826).

Kingma, D. P., & Lei Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations.*

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *IEEE* (pp. 2278–2324).

Li, M., Soltanolkotabi, M., & Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics* (pp. 4313–4324).

Natarajan, N., Dhillon, I. S., Ravikumar, P., & Tewari, A. (2018). Cost-Sensitive Learning with Noisy Labels. *The Journal of Machine Learning Research*, *18*, 1–33.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1944–1952).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Reid, M. D., & Williamson, R. C. (2010). Composite binary losses. *The Journal of Machine Learning Research*, *11*, 2387–2422.

Song, H., Kim, M., Park, D., & Lee, J.-G. (2020). Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. B. (2019). Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics* (pp. 3459–3467).

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., & Bailey, J. (2019). Symmetric Cross Entropy for Robust Learning with Noisy Labels. In *IEEE International Conference on Computer Vision* (pp. 322–330).

Xu, Y., Cao, P., Kong, Y., & Wang, Y. (2019). L$_{DMI}$: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise. In *Advances in Neural Information Processing Systems* (pp. 6222–6233).

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems* (pp. 8778–8788).

Supplementary Material:
Robustness and Reliability When Training With Noisy Labels

A ASYMPTOTIC RISK MINIMISERS

We derive the conditional (or point-wise), asymptotic risk minimisers for categorical cross-entropy (CCE) loss, mean absolute error (MAE) and sigmoid loss in a binary classification setting. The first loss function is strictly proper and the last two are symmetric according to definition 2.5. We also derive the asymptotic risk minimiser for a general symmetric loss. A summary is shown in table 1. Note that for a binary classifier, we assume $f(x) = [f_1(x), 1 - f_1(x)]^\top$.

To derive the risk minimisers, we consider finding the minima of the point-wise risk $\mathcal{J}(f(x)) = \mathbb{E}_{Y|X=x}[\ell(f(x), Y)]$. For convenience, we will sometimes refer to the one-hot version of Y by $e^{(Y)}$. In that case, $e_k^{(Y)}$ refers to the k^{th} element of the vector.

A.1 CCE Loss

Loss function: $\ell(f(x), y) = -\sum_{k=1}^{2} e_k^{(y)} \log f_k(x) = -\log f_y(x)$

Symmetry check:

$$\sum_{k=1}^{K} \ell(f(x), k) = -\sum_{k=1}^{2}\sum_{l=1}^{2} e_l^{(k)} \log f_k(x) = -\sum_{k=1}^{2} \log f_k(x)$$

Not symmetric.

Risk minimiser: $f_1^*(x) = g_1(x)$

Table 1: Some loss functions and their asymptotic risk minimisers in a binary classification setting. We show the predicted probability for class 1, $f_1^*(x)$, but $f_2^*(x) = 1 - f_1^*(x)$. For short notation we use $\gamma \in \text{argmin}_{q\in[0,1]}\ell(q, 1)$ and $\gamma' \in \text{argmax}_{q\in[0,1]}\ell(q, 1)$.

Loss	$f_1^*(x)$	$\sum_{k=1}^{K} \ell(f(x), k)$
CCE	$\mathbb{P}(Y = 1 \mid X = x)$	$-\sum_{k=1}^{2} \log f_k(x)$
MAE	$\mathbb{1}_{\mathbb{P}(Y=1\|X=x)\geq\frac{1}{2}}$	2
Sigmoid	$\mathbb{1}_{\mathbb{P}(Y=1\|X=x)\geq\frac{1}{2}}$	1
Symmetric	$\gamma\mathbb{1}_{\mathbb{P}(Y=1\|X=x)\geq\frac{1}{2}} + \gamma'\mathbb{1}_{\mathbb{P}(Y=1\|X=x)<\frac{1}{2}}$	C

Derivation:

$$\begin{aligned}\mathcal{J}(f(x)) &= -g_1(x)\log f_1(x) - g_2(x)\log f_2(x)\\ &= -g_1(x)\log f_1(x) - (1-g_1(x))\log(1-f_1(x))\end{aligned}$$

$$\Rightarrow \frac{\partial \mathcal{J}(f(x))}{\partial f_1(x)} = -g_1(x)\frac{1}{f_1(x)} + (1-g_1(x))\frac{1}{1-f_1(x)}$$

$$\Rightarrow \frac{\partial \mathcal{J}(f(x))}{\partial f_1(x)} = 0 \Rightarrow f_1^*(x) = g_1(x)$$

A.2 MAE

Loss: $\ell(f(x),y) = \sum_{k=1}^{2}|e_k^{(y)} - f_k(x)|$

Symmetry check:

$$\sum_{k=1}^{K}\ell(f(x),k) = \sum_{k=1}^{2}\sum_{l=1}^{2}|e_l^{(k)} - f_j(x)| = 2\cdot 2 - 2 = 2$$

Symmetric.

Risk minimiser: $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq\frac{1}{2}}$

Derivation:

$$\begin{aligned}\mathcal{J}(f(x)) &= g_1(x)(|1-f_1(x)| + |-f_2(x)|) + g_2(x)(|-f_1(x)| + |1-f_2(x)|)\\ &= 2g_1(x)(1-f_1(x)) + 2(1-g_1(x))f_1(x)\end{aligned}$$

If $g_1(x) \geq \frac{1}{2}$, minimum at $f_1(x) = 1$, otherwise minimum at $f_1(x) = 0$. We can formulate this as $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq\frac{1}{2}}$.

A.3 Sigmoid Loss

Loss function: $\ell(f(x),y) = e_1^{(y)}\frac{1}{1+e^{f_1(x)}} + e_2^{(y)}\frac{e^{f_1(x)}}{1+e^{f_1(x)}}$

Symmetry check:

$$\sum_{k=1}^{2}\ell(f(x),k) = \ell(f(x),1) + \ell(f(x),2) = \frac{1}{1+e^{f_1(x)}} + \frac{e^{f_1(x)}}{1+e^{f_1(x)}} = 1$$

Symmetric.

Risk minimiser: $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq\frac{1}{2}}$

Derivation:

$$\begin{aligned}\mathcal{J}(f(x)) &= g_1(x)\frac{1}{1+e^{f_1(x)}} + g_2(x)\frac{e^{f_1(x)}}{1+e^{f_1(x)}}\\ &= g_1(x)\frac{1}{1+e^{f_1(x)}} + (1-g_1(x))(1-\frac{1}{1+e^{f_1(x)}})\end{aligned}$$

If $g_1(x) \geq \frac{1}{2}$, minimum at $f_1(x) \to \infty$, otherwise minimum at $f_1(x) \to -\infty$. Since $f_1(x) \in [0, 1]$ and $\frac{1}{1+e^{f_1(x)}}$ is a (monotonic) decreasing function in $f_1(x)$, we have a minimum at $f_1(x) = 1$ for $g_1(x) \geq \frac{1}{2}$ and at $f_1(x) = 0$ for $g_1(x) < \frac{1}{2}$. So, $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x) \geq \frac{1}{2}}$.

A.4 General Symmetric Loss

Loss function: $\ell(f(x), y)$

Symmetry check:

$$\sum_{k=1}^{2} \ell(f(x), k) = C$$

Symmetric by definition.

Risk minimiser: $f_1^*(x) = \gamma \mathbb{1}_{\mathbb{P}(Y=1|X=x) \geq \frac{1}{2}} + \gamma' \mathbb{1}_{\mathbb{P}(Y=1|X=x) < \frac{1}{2}}$

Derivation:

$$\begin{aligned}
\mathcal{J}(f(x)) &= g_1(x)\ell(f(x), 1) + g_2(x)\ell(f(x), 2) \\
&= g_1(x)\ell(f(x), 1) + (1 - g_1(x))(C - \ell(f(x), 1)) \\
&= \ell(f(x), 1)(2g_1(x) - 1) + C(1 - g_1(x))
\end{aligned}$$

If $g_1(x) \geq \frac{1}{2}$, minima at $\operatorname{argmin}_{q \in [0,1]} \ell(q, 1)$, otherwise minima at $\operatorname{argmax}_{q \in [0,1]} \ell(q, 1)$. Hence, minima is found at

$$f_1^*(x) = \gamma \mathbb{1}_{\mathbb{P}(Y=1|X=x) \geq \frac{1}{2}} + \gamma' \mathbb{1}_{\mathbb{P}(Y=1|X=x) < \frac{1}{2}}$$

with $\gamma \in \operatorname{argmin}_{q \in [0,1]} \ell(q, 1)$ and $\gamma' \in \operatorname{argmax}_{q \in [0,1]} \ell(q, 1)$. Since we assume $f_1(x) \in [0, 1]$, we must also have $\gamma, \gamma' \in [0, 1]$.

B COMPLETE PROOFS

Complete proofs of the statements, lemmas and propositions from the main paper follow.

Lemma 3.1. *The conditional probability vector $\tilde{g}(x)$ can be written as function of $g(x)$ according to*

$$\tilde{g}(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) g(x) + \frac{\omega(x)}{K-1} \cdot \mathbf{1}_K$$

where $\mathbf{1}_K$ is the vector of ones with length K. Moreover, it holds for any two classes $i, j \in \mathcal{Y}$ that $\tilde{g}_i(x) > \tilde{g}_j(x)$ if and only if $g_i(x) > g_j(x)$.

Proof. For the first part of the proof, note that for any $i \in \mathcal{Y}$, $\sum_{k \neq i} g_k(x) = 1 - g_i(x)$, then

$$\begin{aligned}
\tilde{g}_i(x) &= \sum_{k=1}^{K} \mathbb{P}(\tilde{Y} = i \mid Y = k, X = x) g_k(x) \\
&= (1 - \omega(x))g_i(x) + \frac{\omega(x)}{K-1} \sum_{k \neq i} g_k(x) \\
&= (1 - \omega(x))g_i(x) + \frac{\omega(x)}{K-1}(1 - g_i(x)) \\
&= \left(1 - \frac{\omega(x)K}{K-1}\right) g_i(x) + \frac{\omega(x)}{K-1}.
\end{aligned}$$

Since this holds for all $i \in \mathcal{Y}$, we can write the result on vector form according to

$$\tilde{g}(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) g(x) + \frac{\omega(x)}{K-1} \cdot \mathbf{1}_K$$

where $\mathbf{1}_K$ is a vector of ones with length K. This finishes the first part of the proof.

For the second part of the proof. Take any two classes $i, j \in \mathcal{Y}$ and assume $g_i(x) > g_j(x)$. Then, from the equation derived above and since $0 \leq \omega(x) < \frac{K-1}{K}$, we get

$$\begin{aligned}\tilde{g}_i(x) &= \left(1 - \frac{\omega(x)K}{K-1}\right) g_i(x) + \frac{\omega(x)}{K-1} \\ &> \left(1 - \frac{\omega(x)K}{K-1}\right) g_j(x) + \frac{\omega(x)}{K-1} \\ &= \tilde{g}_j(x).\end{aligned}$$

Hence, $g_i(x) > g_j(x)$ implies $\tilde{g}_i(x) > \tilde{g}_j(x)$.

Next, assume $\tilde{g}_i(x) > \tilde{g}_j(x)$. Then by rearranging the same equation, we obtain the following

$$\begin{aligned}g_i(x) &= \frac{\tilde{g}_i(x) - \frac{\omega(x)}{K-1}}{1 - \frac{\omega(x)K}{K-1}} \\ &> \frac{\tilde{g}_j(x) - \frac{\omega(x)}{K-1}}{1 - \frac{\omega(x)K}{K-1}} \\ &= g_j(x).\end{aligned}$$

Hence, $\tilde{g}_i(x) > \tilde{g}_j(x)$ implies $g_i(x) > g_j(x)$. Therefore, $\tilde{g}_i(x) > \tilde{g}_j(x)$ if and only if $g_i(x) > g_j(x)$. □

Lemma B.1. *For simple non-uniform label noise, $\tilde{g}(x) = g(x)$ if and only if $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$.*

Proof. First, we show that $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$ implies $g(x) = \tilde{g}(x)$. Assume $\omega(x) = 0$, then from lemma 3.1:

$$\tilde{g}(x) = \left(1 - \frac{0 \cdot K}{K-1}\right) g(x) + \frac{0}{K-1} \cdot \mathbf{1}_K = g(x)$$

Similarly, if $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$, then

$$\tilde{g}(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) \frac{1}{K} \cdot \mathbf{1}_K + \frac{\omega(x)}{K-1} \cdot \mathbf{1}_K = \frac{1}{K} \cdot \mathbf{1}_K$$

i.e. $\tilde{g}(x) = g(x)$ and so, $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$ implies $g(x) = \tilde{g}(x)$.

Next, we show that $\tilde{g}(x) = g(x)$ implies $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$. Assume $\tilde{g}(x) = g(x)$, then from lemma 3.1

$$\tilde{g}(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) g(x) + \frac{\omega(x)}{K-1} \cdot \mathbf{1}_K = g(x)$$

which implies $\omega(x) = 0$ or

$$Kg(x) - \mathbf{1}_K = 0$$

or, equivalently, $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$. Therefore, if $\tilde{g}(x) = g(x)$, then $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$. Hence, we have shown that $\tilde{g}(x) = g(x)$ if and only if $\omega(x) = 0$ or $g(x) = \frac{1}{K} \cdot \mathbf{1}_K$. □

Proposition 3.1. *Assume that the prediction is taken as the most probable class, then $\tilde{g}$ has the same decision boundaries as g.*

Proof. To show that $\tilde{g}$ and g have the same decision boundaries, we will show that $\tilde{g}_i(x) = \tilde{g}_j(x)$ if and only if $g_i(x) = g_j(x)$ for all $x \in \mathcal{X}$ for which this holds and for any two classes $i, j \in \mathcal{Y}$. In addition, lemma 3.1 states that classes are not flipped, as $\tilde{g}_i(x) > \tilde{g}_j(x)$ if and only if $g_i(x) > g_j(x)$.

We first show that if $g_i(x) = g_j(x)$, then $\tilde{g}_i(x) = \tilde{g}_j(x)$. For $i, j \in \mathcal{Y},\ i \neq j$, assume $g_i(x) = g_j(x)$. Then, for any x for which this holds (i.e. any x along the decision boundary between classes i and j)

$$\begin{aligned}\tilde{g}_j(x) &= \left(1 - \frac{\omega(x)K}{K-1}\right) g_j(x) + \frac{\omega(x)}{K-1}\\ &= \left(1 - \frac{\omega(x)K}{K-1}\right) g_i(x) + \frac{\omega(x)}{K-1}\\ &= \tilde{g}_i(x).\end{aligned}$$

This follows directly from lemma 3.1 and since $g_i(x) = g_j(x)$.

Next, we show that $\tilde{g}_i(x) = \tilde{g}_j(x)$ implies $g_i(x) = g_j(x)$. Assume $\tilde{g}_i(x) = \tilde{g}_j(x)$ for some $i, j \in \mathcal{Y},\ i \neq j$. From lemma 3.1, we get

$$g_i(x) = \frac{\tilde{g}_i(x) - \frac{\omega(x)}{K-1}}{1 - \frac{\omega(x)K}{K-1}}$$

and therefore,

$$\begin{aligned}g_i(x) &= \frac{\tilde{g}_i(x) - \frac{\omega(x)}{K-1}}{1 - \frac{\omega(x)K}{K-1}}\\ &= \frac{\tilde{g}_j(x) - \frac{\omega(x)}{K-1}}{1 - \frac{\omega(x)K}{K-1}}\\ &= g_j(x),\end{aligned}$$

i.e. if $\tilde{g}_i(x) = \tilde{g}_j(x)$, then $g_i(x) = g_j(x)$.

This shows that for any $x \in \mathcal{X}$ it holds that $\tilde{g}_i(x) = \tilde{g}_j(x)$ if and only if $g_i(x) = g_j(x)$. Note that the argument easily extends to equality between several classes, i.e. $\tilde{g}_i(x) = \tilde{g}_j(x) = \cdots = \tilde{g}_l(x)$ if and only if $g_i(x) = g_j(x) = \cdots = g_l(x)$, since we still consider pairwise equality. In turn, this implies, together with lemma 3.1, that $\tilde{g}$ and g share decision boundaries.

□

Proposition 3.2. *The average conditional entropy of $\tilde{g}$ is higher than that of g, that is $\mathbb{E}_X[\mathcal{H}[\tilde{g}(X)]] > \mathbb{E}_X[\mathcal{H}[g(X)]]$.*

Proof. To show that $\tilde{g}$ has a higher average conditional entropy than g, we first note that the conditional entropy is a strictly concave function. This follows from the fact that the entropy is a sum of strictly concave functions $h(z) = -z\log(z)$. Hence, for any two distinct vectors $f^1(x), f^2(x)$

$$\mathcal{H}[f^\lambda(x)] > (1-\lambda)\mathcal{H}[f^1(x)] + \lambda\mathcal{H}[f^2(x)], \quad \lambda \in (0,1)$$

with $f^\lambda(x) = (1-\lambda)f^1(x) + \lambda f^2(x)$. The unique maximum of the conditional entropy is found at $u = \frac{1}{K} \cdot \mathbf{1}_K$, the vector of uniform probability.

Notice that for $\lambda = \frac{\omega(x)K}{K-1}$, we can use lemma 3.1 to write $\tilde{g}(x)$ as a linear combination of the form

$$\tilde{g}(x) = (1-\lambda)g(x) + \lambda\frac{1}{K} \cdot \mathbf{1}_K$$

Let $f^1(x) = g(x)$, $f^2(x) = u$ (the uniform vector) and $\lambda = \frac{\omega(x)K}{K-1}$. For $\omega(x) > 0$ and $g(x) \neq u$, we have

$$\begin{aligned}\mathcal{H}[\tilde{g}(x)] &> (1-\lambda)\mathcal{H}[g(x)] + \lambda\mathcal{H}[u] \\ &> (1-\lambda)\mathcal{H}[g(x)] + \lambda\mathcal{H}[g(x)] \\ &= \mathcal{H}[g(x)].\end{aligned}$$

The last inequality follows since the maximum at u is unique and since $\lambda > 0$. Hence, for $g(x) \neq u$ and $\omega(x) > 0$, it holds that $\mathcal{H}[\tilde{g}(x)] > \mathcal{H}[g(x)]$. For $g(x) = u$ or if $\omega(x) = 0$, it follows from lemma B.1 that $\tilde{g}(x) = g(x)$ and, therefore, $\mathcal{H}[\tilde{g}(x)] = \mathcal{H}[g(x)]$.

Let $\mathcal{X}_1 = \{x \in \mathcal{X};\ \tilde{g}(x) = g(x)\}$ and $\mathcal{X}_2 = \{x \in \mathcal{X};\ \tilde{g}(x) \neq g(x)\}$. Moreover, let μ_X be the marginal probability distribution of X. Since $\mathbb{P}\left(\{\omega(X) > 0\} \cap \{g(X) \neq u\}\right) > 0$ implies $\mathbb{P}(X \in \mathcal{X}_2) > 0$ (lemma B.1), we have

$$\begin{aligned}\mathbb{E}_X[\mathcal{H}[\tilde{g}(X)]] &= \int_{\mathcal{X}_1} \mathcal{H}[\tilde{g}(x)]\mu_X(dx) + \int_{\mathcal{X}_2} \mathcal{H}[\tilde{g}(x)]\mu_X(dx) \\ &> \int_{\mathcal{X}_1} \mathcal{H}[g(x)]\mu_X(dx) + \int_{\mathcal{X}_2} \mathcal{H}[g(x)]\mu_X(dx) \\ &= \mathbb{E}_X[\mathcal{H}[g(X)]]\end{aligned}$$

meaning that $\tilde{g}$ has a higher conditional entropy than g on average over $\mathcal{X}$.

□

Lemma B.2. *Let $M \in \underset{k\in\mathcal{Y}}{argmax}\, \tilde{g}_k(x)$, then $g_M(x) \geq \tilde{g}_M(x)$ with equality if and only if $g(x) = \tilde{g}(x)$.*

Proof. We first show that for any $M \in \underset{k}{\operatorname{argmax}}\, \tilde{g}_k(x)$, it holds that $g_M(x) \geq \tilde{g}_M(x)$. Note that it follows from lemma 3.1 and proposition 3.1 that

$$\underset{k}{\operatorname{argmax}}\, g_k(x) = \underset{k}{\operatorname{argmax}}\, \tilde{g}_k(x)$$

such that if $M \in \underset{k}{\operatorname{argmax}}\, \tilde{g}_k(x)$, then $M \in \underset{k}{\operatorname{argmax}}\, g_k(x)$. Note also that it must hold that $\tilde{g}_M(x), g_M(x) \geq \frac{1}{K}$, since $\tilde{g}(x)$ and $g(x)$ are both vectors of norm 1. From lemma 3.1, we have

$$\begin{aligned}\tilde{g}_M(x) &= \left(1 - \frac{\omega(x)K}{K-1}\right) g_M(x) + \frac{\omega(x)}{K-1} \\ &= (1 - Kg_M(x))\frac{\omega(x)}{K-1} + g_M(x)\end{aligned}$$

which is a linear function in $\omega(x)$. Since $g_M(x) \geq \frac{1}{K}$, $\tilde{g}_M(x)$ decreases with $\omega(x)$ and an upper bound is found at $\omega(x) = 0$

$$\tilde{g}_M(x) \leq (1 - Kg_M(x))\frac{\omega(x)}{K-1} + g_M(x)\bigg|_{\omega(x)=0} = g_M(x),$$

which finishes the first part of the proof.

Next, we show that $g_M(x) = \tilde{g}_M(x)$, for any $M \in \underset{k}{\operatorname{argmax}}\, g_k(x)$, if and only if $g(x) = \tilde{g}(x)$. First, it follows directly that if $g(x) = \tilde{g}(x)$ then $g_k(x) = \tilde{g}_k(x)\ \forall k \in \mathcal{Y}$ and therefore, $g_M(x) = \tilde{g}_M(x)$. Second, to see that $g_M(x) = \tilde{g}_M(x)$ implies $g(x) = \tilde{g}(x)$, we use lemma 3.1 with $g_M(x) = \tilde{g}_M(x)$, to get

$$\tilde{g}_M(x) = \left(1 - \frac{\omega(x)K}{K-1}\right) \tilde{g}_M(x) + \frac{\omega(x)}{K-1}$$

which implies $\omega(x) = 0$ or

$$(K\tilde{g}_M(x) - 1) = 0,$$

i.e. $\tilde{g}_M(x) = \frac{1}{K}$. For $\omega(x) = 0$, it follows directly from lemma 3.1 that $g(x) = \tilde{g}(x)$. In the second case, notice that $\tilde{g}_M(x) = \frac{1}{K}$ must imply that $\tilde{g}_k = \frac{1}{K}\ \forall k \in \mathcal{Y}$, since $M \in \underset{k}{\operatorname{argmax}}\ \tilde{g}_k(x)$ and $\|\tilde{g}_k(x)\|_1 = 1$. Then, from lemma 3.1, we get

$$\begin{aligned} g(x) &= \frac{(\frac{1}{K} - \frac{\omega(x)}{K-1}) \cdot \mathbf{1}_K}{1 - \frac{\omega(x)K}{K-1}} \\ &= \frac{\frac{1}{K}(1 - \frac{\omega(x)K)}{K-1}}{(\frac{1}{K} - \frac{\omega(x)K}{K-1})} \\ &= \frac{1}{K} \cdot \mathbf{1}_K \end{aligned}$$

which means that $g(x) = \tilde{g}(x)$. Therefore, $g_M(x) \geq \tilde{g}_M(x)$ with equality if and only if $g(x) = \tilde{g}(x)$.

□

Proposition 3.3. *The vector of conditional probabilities $\tilde{g}(X)$ is not calibrated with respect to the distribution $\mathbb{P}(Y \mid X)$ over clean labels.*

Proof. Let $Z = g(X)$ and $\tilde{Z} = \tilde{g}(X)$. For $\tilde{g}(X)$ to be calibrated we require,

$$\mathbb{P}(Y = k \mid \tilde{Z}) = \tilde{Z}_k,\ \forall k \in \mathcal{Y}$$

almost surely. Hence, to show that $\tilde{g}$ is not calibrated, it is enough to show that $\mathbb{P}(Y = k \mid \tilde{Z}) \neq \tilde{Z}_k$, with probability larger than 0, for any $k \in \mathcal{Y}$. Let

$$M = \inf \underset{k}{\operatorname{argmax}}\ \tilde{Z}_k,$$

where the infimum is taken just to select a unique index in the case when the maximising argument is not unique. We know from lemma 3.1 and proposition 3.1 that

$$M = \inf \underset{k}{\operatorname{argmax}}\ Z_k$$

meaning that M is a $\sigma(\tilde{Z}) \cap \sigma(Z)$-measurable random variable. From lemma B.2 we know that

$$Z_M \geq \tilde{Z}_M$$

with equality if and only if $Z = \tilde{Z}$ almost surely. From the tower property of conditional expectation we get

$$\mathbb{P}(Y = M \mid \tilde{Z}) = \mathbb{E}_{Z|\tilde{Z}}[\mathbb{E}_{Y|Z,\tilde{Z}}[\mathbb{1}_{Y=M}]] = \mathbb{E}_{Z|\tilde{Z}}[Z_M]$$

where the second equality follows from the fact that Y is conditionally independent of $\tilde{Z}$ given Z, which in turn follows from the properties of the selected noise model, and that

$$\mathbb{P}(Y = M \mid Z) = Z_M$$

by definition. Hence, since $\mathbb{P}(Z \neq \tilde{Z}) > 0$ by assumption, there is a set with measure strictly larger than zero on which

$$\mathbb{P}(Y = M \mid \tilde{Z}) > \tilde{Z}_M$$

and $\tilde{g}(X)$ is not calibrated for $\mathbb{P}(Y \mid X)$.

□

Proposition 3.4. *Robust loss functions (definition 2.4) are not strictly proper.*

Proof. Assume that the loss function ℓ is robust according to definition 2.4, i.e. $\ell \in \mathcal{L}_{\mathrm{R}}$. Since ℓ is robust, it holds for all $f^* \in \underset{f}{\operatorname{argmin}}\, \mathcal{R}_\ell(f)$ that

$$\tilde{\mathcal{R}}_\ell(f^*) \leq \tilde{\mathcal{R}}_\ell(f), \quad \forall f \in \mathcal{F},$$

with equality only if f is also in the set of asymptotic risk minimisers.

Assume now that ℓ is strictly proper. Then,

$$\underset{f}{\operatorname{argmin}}\, \mathcal{R}_\ell(f) = \{g\}.$$

In parallel, strictly properness implies

$$\underset{f}{\operatorname{argmin}}\, \tilde{\mathcal{R}}_\ell(f) = \{\tilde{g}\}$$

and therefore,

$$\tilde{\mathcal{R}}_\ell(\tilde{g}) \leq \tilde{\mathcal{R}}_\ell(g),$$

with equality only in the noise-free case where $\tilde{g} = g$. Otherwise, this is a contradiction. As a result, $\ell \notin \mathcal{L}_{\mathrm{SP}}$. Since ℓ is an arbitrary robust loss function, we have shown that

$$\mathcal{L}_{\mathrm{R}} \cap \mathcal{L}_{\mathrm{SP}} = \emptyset.$$

□

Proposition 3.5. *Symmetric loss functions (definition 2.5) are not calibration-based strictly proper.*

Proof. For $\mathcal{L}_{\mathrm{S}} \cap \mathcal{L}_{\mathrm{CSP}} = \emptyset$ to hold, we require that every symmetric loss function has at least one asymptotic risk minimiser that is not calibrated. That is, every $\ell \in \mathcal{L}_{\mathrm{S}}$, has at least one risk minimiser $f^* \notin \mathcal{F}_{\mathcal{C}}$, for at least one conditional distribution $\mathbb{P}(Y \mid X)$ over the target, Y, and one probability distribution, μ_X, over the input, X.

Consider the binary case and a symmetric loss function with (point-wise) asymptotic risk minimisers defined in eq. (4). For an arbitrary symmetric loss function $\ell \in \mathcal{L}_{\mathrm{S}}$ and if at least one risk minimiser exists, one of the following holds for the parameters γ, γ' in eq. (4):

(i) Both γ and γ' are unique.

(ii) At least one of γ, γ' is not unique.

Following the definitions of γ and γ', and since ℓ is solely a function of a probability vector q and a label y, we know that γ and γ' are both independent of the given input x. Hence, for (i), the full risk minimiser will take the form

$$f_1^*(x) = \begin{cases} \gamma, & \text{if } \mathbb{P}(Y = 1 \mid X = x) \geq \frac{1}{2} \\ \gamma', & \text{otherwise} \end{cases}$$

$\forall x \in \mathcal{X}$ and where γ, γ' are constants.

Let $\mathcal{X}_1 = \{x \in \mathcal{X};\ \mathbb{P}(Y = 1 \mid X = x) \geq \frac{1}{2}\}$ and $\mathcal{X}_2 = \{x \in \mathcal{X};\ \mathbb{P}(Y = 1 \mid X = x) < \frac{1}{2}\}$. For f^* to be calibrated, we require

$$\mathbb{P}(Y = 1 \mid f_1^*(X) = \gamma) = \mathbb{P}(Y = 1 \mid X \in \mathcal{X}_1) = \gamma,$$
$$\mathbb{P}(Y = 1 \mid f_1^*(X) = \gamma') = \mathbb{P}(Y = 1 \mid X \in \mathcal{X}_2) = \gamma'.$$

This is true if γ (coincidentally) matches the average probability of $Y = 1$ over $\mathcal{X}_1$ and γ' is equal to the average probability of $Y = 1$ over $\mathcal{X}_2$ according to

$$\mathbb{P}(Y = 1 \mid X \in \mathcal{X}_1) = \frac{\int_{\mathcal{X}_1} \mathbb{P}(Y = 1 \mid X = x)\mu_X(dx)}{\int_{\mathcal{X}_1} \mu_X(dx)} = \gamma,$$

$$\mathbb{P}(Y = 1 \mid X \in \mathcal{X}_2) = \frac{\int_{\mathcal{X}_2} \mathbb{P}(Y = 1 \mid X = x)\mu_X(dx)}{\int_{\mathcal{X}_2} \mu_X(dx)} = \gamma'.$$

To see that the model f^* is not calibrated in general, assume that there is a conditional probability $\mathbb{P}(Y \mid X)$ and marginal distribution μ_X for which the model is calibrated, i.e. the equations above hold. Then, consider any other data distribution for which μ_X is the same but where the conditional probability over Y can be described by

$$\mathbb{P}'(Y = 1 \mid X = x) = (1 - 2\alpha)\mathbb{P}(Y = 1 \mid X = x) + \alpha, \quad 0 < \alpha < \frac{1}{2},$$

such that $\mathcal{X}_1$ and $\mathcal{X}_2$, and consequently f^*, remain the same. For this data distribution and for any suitable solution $\gamma \in [\frac{1}{2}, 1]$, $\gamma' \in [0, \frac{1}{2})$, we find that

$$\begin{aligned}\mathbb{P}'(Y = 1 \mid X \in \mathcal{X}_1) &= \frac{\int_{\mathcal{X}_1} \mathbb{P}'(Y = 1 \mid X = x)\mu_X(dx)}{\int_{\mathcal{X}_1} \mu_X(dx)} \\ &= \frac{\int_{\mathcal{X}_1} \left((1 - 2\alpha)\mathbb{P}(Y = 1 \mid X = x) + \alpha\right)\mu_X(dx)}{\int_{\mathcal{X}_1} \mu_X(dx)} \\ &= (1 - 2\alpha)\gamma + \alpha \leq \gamma\end{aligned}$$

with equality only if $\gamma = \frac{1}{2}$. Similarly,

$$\begin{aligned}\mathbb{P}'(Y = 1 \mid X \in \mathcal{X}_2) &= \frac{\int_{\mathcal{X}_2} \mathbb{P}'(Y = 1 \mid X = x)\mu_X(dx)}{\int_{\mathcal{X}_2} \mu_X(dx)} \\ &= (1 - 2\alpha)\gamma' + \alpha > \gamma'.\end{aligned}$$

As a result, f^* is not calibrated for the conditional distribution $\mathbb{P}'(Y \mid X)$. Consequently, we can conclude that there exists a conditional distribution $\mathbb{P}(Y \mid X)$ and marginal μ_X for which the asymptotic risk minimiser f^* of ℓ is not calibrated.

Now, assume (ii) holds, i.e. the loss ℓ has two or more (point-wise) asymptotic risk minimers. Since every combination (γ, γ'), from the set of point-wise asymptotic risk minimisers of ℓ, is equally viable for every $x \in \mathcal{X}$ (they are all independent of x and minimise the point-wise risk), at least one risk minimiser can be formulated according to the unique risk minimiser in case (i). Hence, there exist at least one risk minimiser f^* for which it holds that $f^* \notin \mathcal{F}_{\mathcal{C}}$ for some $\mathbb{P}(Y \mid X)$ and μ_X.

With the arguments put forth, it holds for all symmetric loss functions that for some conditional distribution $\mathbb{P}(Y \mid X)$ and marginal μ_X, there exists an asymptotic risk minimiser that is not calibrated. It follows that symmetric loss functions are not calibration-based strictly proper, i.e. $\mathcal{L}_{\text{S}} \cap \mathcal{L}_{\text{CSP}} = \emptyset$. □

C EXPERIMENTAL DETAILS

The empirical experiments were performed using Python and the Pytorch deep learning library (Paszke et al., 2019). Experimental details follow in this section.

C.1 Simple Noise Example

For the simple noise example in fig. 2, the models used were fully connected neural networks with one hidden layer of 50 hidden units and with ReLU activation. The models were trained on two-dimensional circle data[2] with

[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_circles.html

5,000 observations, generated with the scikit-learn library (Pedregosa et al., 2011). For the noisy data set, labels were flipped with a uniform flip probability of $\omega = 0.2$. Both models were trained with categorical cross-entropy loss and ADAM optimization (Kingma & Lei Ba, 2015). We used a constant learning rate of 0.1 and a batch size of 100. A separate validation set of size 1,000 was used for early stopping, where the training was stopped if the current validation loss value exceeded the minimum achieved with more than 10%. From that, the model with the smallest validation loss was selected. The models were evaluated in terms of accuracy on a separately generated, clean, test data set of 1,000 samples. In addition, plots of the predicted class 1 probability for each model were generated on a grid of range $(-1.5, 1.5)$ in both dimensions.

C.2 Robust Loss Functions and Overfitting

For the empirical evaluation of robustness against overfitting, all models used were fully connected neural networks with one hidden layer of 500 hidden units and LeakyReLU activation. The models were trained with ADAM optimisation (Kingma & Lei Ba, 2015) with a constant learning rate of 0.005 and a batch size of 100. The hyperparameters were selected such that the general trends of the training dynamics could be observed.

The data used for training was the MNIST training data set (LeCun, Bottou, Bengio, & Haffner, 1998) where 50,000 of the data points were used for training and 10,000 for validation. The images were flattened prior to training. Two sets of data with symmetric label noise were created by random flipping of labels. For the first noisy set, a flip probability of $\omega = 0.3$ was used and for the second set, we used $\omega = 0.5$. The original data set was assumed to be noise free, corresponding to a flip probability of $\omega = 0.0$.

The models trained with mean absolute error (we use the built-in L1Loss in Pytorch with mean reduction, which in our framework corresponds to training with MAE $\cdot \frac{1}{K}$) were trained for 5,000 epochs and evaluated every 10^{th} epoch. The models were initialised randomly or from the weights obtained by a separate model trained with categorical cross-entropy (CCE) loss on the corresponding clean or noisy data set. The models trained with CCE loss, both for the purpose of pre-training and for the purpose of separate evaluation, were trained for 500 epochs and evaluated every epoch, if relevant.

The models were evaluated both on the data set on which they were trained and on the separate, presumably clean, MNIST test data set of 10,000 observations. We evaluated all models in terms of accuracy. In addition, we compared the training loss (MAE) on the respective training sets for the models trained with MAE, with and without pre-training.

D INFORMATION-THEORETIC LOSS FUNCTION

Apart from the symmetric loss functions identified in Ghosh, Kumar, and Sastry, 2017; Ghosh, Manwani, and Sastry, 2015, the information-theoretic loss function proposed by (Xu, Cao, Kong, & Wang, 2019) is robust according to definition 2.4 under symmetric label noise. We will show that while this loss function is robust, it does not recover g and it is not calibration-based strictly proper. The information-theoretic loss function is based on Determinant-based Mutual Information (DMI) and is defined as

$$\ell(f(X), Y) = -\log |\det(\mathbb{P}(\hat{Y}, Y)|, \quad \hat{Y} \sim f(X).$$

In the equation, we use $|\cdot|$ to denote the absolute value and $\mathbb{P}(\hat{Y}, Y)$ should be interpreted as the $K \times K$ probability matrix corresponding to the joint distribution of $\hat{Y}$ and Y.

For instance-independent label noise, e.g. symmetric noise, it is possible to show (Xu et al., 2019) that

$$\ell(f(X), \tilde{Y}) = \ell(f(X), Y) + C$$

for a constant C. Following this, it can be concluded that definition 2.4 is fulfilled.

Next, we derive the asymptotic (risk) minimisers of the information-theoretic loss function. Assume $\mathcal{Y} = \{1, 2\}$, then

$$\ell(f(X), Y) = -\log\ |\mathbb{P}(\hat{Y} = 1, Y = 1)\mathbb{P}(\hat{Y} = 2, Y = 2) - \mathbb{P}(\hat{Y} = 1, Y = 2)\mathbb{P}(\hat{Y} = 2, Y = 1)|.$$

The loss function is minimised when the absolute value of the determinant is maximised. To determine what this means for the model f, we factorise $\mathbb{P}(\hat{Y}, Y)$ according to $\mathbb{P}(\hat{Y}, Y) = \mathbb{P}(\hat{Y} \mid Y)\mathbb{P}(Y)$ and use $\mathbb{P}(\hat{Y} = 2 \mid Y = i) = 1 - \mathbb{P}(\hat{Y} = 1 \mid Y = i)$ to obtain

$$\begin{aligned}\ell(f(X), Y) &= -\log\ |\mathbb{P}(Y = 1)(1 - \mathbb{P}(Y = 1))(\mathbb{P}(\hat{Y} = 1 \mid Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 2))| \\ &= -\log\ |\mathbb{P}(Y = 1)(1 - \mathbb{P}(Y = 1))(\mathbb{E}_{\hat{Y}|Y=1}[\mathbb{1}_{\hat{Y}=1}] - \mathbb{E}_{\hat{Y}|Y=2}[\mathbb{1}_{\hat{Y}=1}])|.\end{aligned}$$

Using the tower property of conditional expectation, we get

$$\begin{aligned}\ell(f(X), Y) &= -\log\ |\mathbb{P}(Y = 1)(1 - \mathbb{P}(Y = 1))(\mathbb{E}_{X|Y=1}[\mathbb{E}_{\hat{Y}|X,Y=1}[\mathbb{1}_{\hat{Y}=1}]] - \mathbb{E}_{X|Y=2}[\mathbb{E}_{\hat{Y}|X,Y=2}[\mathbb{1}_{\hat{Y}=1}]])| \\ &= -\log\ |\mathbb{P}(Y = 1)(1 - \mathbb{P}(Y = 1))(\mathbb{E}_{X|Y=1}[f_1(X)] - \mathbb{E}_{X|Y=2}[f_1(X)])|,\end{aligned}$$

where the last equality follows as $\hat{Y}$ is independent of Y given X. Next, we use Bayes' theorem to rewrite the expression further

$$\begin{aligned}\ell(f(X), Y) &= -\log\ |(1 - \mathbb{P}(Y = 1))\mathbb{E}_X[f_1(X)\mathbb{P}(Y = 1 \mid X)] - \mathbb{P}(Y = 1)\mathbb{E}_X[f_1(X)(1 - \mathbb{P}(Y = 1|X))]| \\ &= -\log\ |\mathbb{E}_X[f_1(X)(\mathbb{P}(Y = 1|X) - \mathbb{P}(Y = 1))]|.\end{aligned}$$

As $0 \leq f_1(X) \leq 1$, the loss is minimised if $f_1(x) = 1$ (or, alternatively, $f_1(x) = 0$) for all $x \in \mathcal{X}$ for which $\mathbb{P}(Y = 1 \mid X = x) - \mathbb{P}(Y = 1) \geq 0$ and $f_1(x) = 0$ ($f_1(x) = 1$) for all $x \in \mathcal{X}$ with $\mathbb{P}(Y = 1 \mid X = x) - \mathbb{P}(Y = 1) < 0$. Hence, the minima are found at $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq\mathbb{P}(Y=1)}$ and $f_1^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)<\mathbb{P}(Y=1)}$. Notice that for balanced classes, i.e. $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$, the first minimiser is the same as the risk minimisers for e.g. MAE and Sigmoid loss. Evidently, the information-theoretic loss function does not recover g. In addition, just as with the minima of symmetric loss functions (see proof of proposition 3.5), the asymptotic minima of the information-theoretic loss function are not calibrated in general.

References

Ghosh, A., Kumar, H., & Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence* (pp. 1919–1925).

Ghosh, A., Manwani, N., & Sastry, P. S. (2015). Making risk minimization tolerant to label noise. *Neurocomputing*, *160*, 93–107.

Kingma, D. P., & Lei Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations.*

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *IEEE* (pp. 2278–2324).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* (pp. 8024–8035).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Xu, Y., Cao, P., Kong, Y., & Wang, Y. (2019). L$_{DMI}$: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise. In *Advances in Neural Information Processing Systems* (pp. 6222–6233).